ADVANCES IN SOFT COMPUTING 49

Juan M. Corchado
Juan F. De Paz
Miguel P. Rocha
Florentino Fernández Riverola (Eds.)

2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics (IWPACBB 2008)

Advances in Soft Computing 49

Editor-in-Chief: J. Kacprzyk

Advances in Soft Computing

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Abraham Ajith, Bernard de Bacts, Mario Köppen, Bertram Nickolay (Eds.)
Applied Soft Computing Technologies: The Challenge of Complexity, 2006
ISBN 978-3-540-31649-7

Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, Kryzysztof Trojanowski (Eds.)
Intelligent Information Processing and Web Mining, 2006
ISBN 978-3-540-33520-7

Ashutosh Tiwari, Joshua Knowles, Erel Avineri, Keshav Dahal, Rajkumar Roy (Eds.)
Applications and Soft Computing, 2006
ISBN 978-3-540-29123-7

Bernd Reusch, (Ed.)
Computational Intelligence, Theory and Applications, 2006
ISBN 978-3-540-34780-4

Jonathan Lawry, Enrique Miranda, Alberto Bugarín Shoumei Li, María Á. Gil, Przemysław Grzegorzewski, Olgierd Hryniewicz,
Soft Methods for Integrated Uncertainty Modelling, 2006
ISBN 978-3-540-34776-7

Ashraf Saad, Erel Avineri, Keshav Dahal, Muhammad Sarfraz, Rajkumar Roy (Eds.)
Soft Computing in Industrial Applications, 2007
ISBN 978-3-540-70704-2

Bing-Yuan Cao (Ed.)
Fuzzy Information and Engineering, 2007
ISBN 978-3-540-71440-8

Patricia Melin, Oscar Castillo, Eduardo Gómez Ramírez, Janusz Kacprzyk, Witold Pedrycz (Eds.)
Analysis and Design of Intelligent Systems Using Soft Computing Techniques, 2007
ISBN 978-3-540-72431-5

Oscar Castillo, Patricia Melin, Oscar Montiel Ross, Roberto Sepúlveda Cruz, Witold Pedrycz, Janusz Kacprzyk (Eds.)
Theoretical Advances and Applications of Fuzzy Logic and Soft Computing, 2007
ISBN 978-3-540-72433-9

Katarzyna M. Węgrzyn-Wolska, Piotr S. Szczepaniak (Eds.)
Advances in Intelligent Web Mastering, 2007
ISBN 978-3-540-72574-9

Emilio Corchado, Juan M. Corchado, Ajith Abraham (Eds.)
Innovations in Hybrid Intelligent Systems, 2007
ISBN 978-3-540-74971-4

Marek Kurzynski, Edward Puchala, Michal Wozniak, Andrzej Zolnierek (Eds.)
Computer Recognition Systems 2, 2007
ISBN 978-3-540-75174-8

Van-Nam Huynh, Yoshiteru Nakamori, Hiroakira Ono, Jonathan Lawry, Vladik Kreinovich, Hung T. Nguyen (Eds.)
Interval / Probabilistic Uncertainty and Non-classical Logics, 2008
ISBN 978-3-540-77663-5

Ewa Pietka, Jacek Kawa (Eds.)
Information Technologies in Biomedicine, 2008
ISBN 978-3-540-68167-0

Didier Dubois, M. Asunción Lubiano, Henri Prade, María Ángeles Gil, Przemysław Grzegorzewski, Olgierd Hryniewicz (Eds.)
Soft Methods for Handling Variability and Imprecision, 2008
ISBN 978-3-540-85026-7

Juan M. Corchado, Juan F. De Paz, Miguel P. Rocha, Florentino Fernández Riverola (Eds.)
2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics (IWPACBB 2008), 2009
ISBN 978-3-540-85860-7

Juan M. Corchado,
Juan F. De Paz, Miguel P. Rocha,
Florentino Fernández Riverola (Eds.)

2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics (IWPACBB 2008)

Springer

Editors

Juan M. Corchado
Departamento de Informática y Automática
Facultad de Ciencias
Universidad de Salamanca
Plaza de la Merced S/N
37008, Salamanca
Spain
E-mail: corchado@usal.es

Juan F. De Paz
Departamento de Informática y Automática
Facultad de Ciencias
Universidad de Salamanca
Plaza de la Merced S/N
37008, Salamanca
Spain
E-mail: fcofds@usal.es

Miguel P. Rocha
Informática/CCTC
Universidade do Minho
Campus de Gualtar
4710-057 Braga
Portugal
E-mail: mrocha@di.uminho.pt

Florentino Fernández Riverola
E.S.E.I.: Escuela Superior de Ingeniería Informática
Departamento de Informática
Universidad de Vigo
Edificio Politécnico. Despacho 408.
Campus Universitario As Lagoas s/n.
32004 - Ourense
Spain
E-mail: riverola@ei.uvigo.es

ISBN 978-3-540-85860-7 e-ISBN 978-3-540-85861-4

DOI 10.1007/978-3-540-85861-4

Advances in Soft Computing ISSN 1615-3871

Library of Congress Control Number: 2008933601

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

5 4 3 2 1 0

springer.com

Preface

The success of Bioinformatics in recent years has been prompted by research in molecular biology and medicine in initiatives like the human genome project. The volume and diversification of data has increased so much that it is very hard if not impossible to analyze it by human experts.

The analysis of this growing body of data, intensified by the development of a number of high-throughput experimental techniques that are generating the so called 'omics' data, has prompted for new computational methods. New global approaches, such as Systems Biology, have been emerging replacing the reductionist view that dominated biology research in the last decades, requiring the coordinated efforts of biological researchers with those related to data analysis, mathematical modelling and computer science. Computational methods have been helping in tasks related to knowledge discovery, modelling and optimization tasks.

This workshop brings the opportunity to discuss applications of Bioinformatics and Computational Biology exploring the interactions between computer scientists, biologists and other scientific researchers. The IWPACBB technical program includes 29 papers (23 long papers and 6 short papers) selected from a submission pool of 51 papers, from 9 different countries.

We thank the excellent work of the local organization members and also from the members of the Program Committee for their excellent reviewing work.

October 2008

Juan M. Corchado
Juan F. De Paz
Miguel P. Rocha
Florentino Fernández Riverola

Organization

General Co-chairs

Juan M. Corchado – University of Salamanca (Spain)
Juan F. De Paz – University of Salamanca (Spain)
Miguel P. Rocha – CCTC, Univ. Minho (Portugal)
Florentino Fernández – University of Vigo (Spain)

Program Committee

Miguel P. Rocha (Chairman) – CCTC, Univ. Minho (Portugal)
Florentino Fernández (Cochairman) – University of Vigo (Spain)
Alicia Troncoso – Universidad of Pablo de Olavide (Spain)
Álvaro Herrero – Universidad of Burgos (Spain)
Anália Lourenço – University of Minho (Portugal)
Arlindo Oliveira – INESC-ID, Lisboa (Portugal)
B. Cristina Pelayo – University of Oviedo (Spain)
Bruno Baruque – Universidad de Burgos (Spain)
Carmen Tejedor – University of Salamanca (Spain)
Daniel Gayo – University of Oviedo (Spain)
Emilio S. Corchado – Universidad de Burgos (Spain)
Eugénio C. Ferreira – IBB/CEB, University of Minho (Portugal)
Fernando Diaz-Gómez – University of Valladolid (Spain)
Isabel C. Rocha – IBB/CEB, University of Minho (Portugal)
Jacob Koehler – University of Tromsø (Norway)
Jesús M. Hernández – University of Salamanca (Spain)
Jorge Alípio – LIAAD/INESC, Porto LA (Portugal)
Jorge Vieira – IBMC, Porto (Portugal)
José Adserias – University of Salamanca (Spain)
José L. López – University of Salamanca (Spain)
Juan M. Cueva – University of Oviedo (Spain)
Júlio R. Banga – IIM/CSIC, Vigo (Spain)
Kiran R. Patil – Biocentrum, DTU (Portugal)
Lourdes Borrajo – Universidad de Vigo (Spain)
Luis M. Rocha – Indiana University (USA)
Margarida Casal – University of Minho (Portugal)

Maria J. Ramos – FCUP, University of Porto (Portugal)
Nuno Fonseca – IBMC, Porto (Portugal)
Oscar Sanjuan – University of Oviedo (Spain)
Paulo Azevedo – University of Minho (Portugal)
Paulino Gómez-Puertas – University Autónoma de Madrid (Spain)
Pierre Baldi – University of California Irvine (USA)
Rui Camacho – LIACC/FEUP, University of Porto (Portugal)
Rui Brito – University of Coimbra (Portugal)
Rui C. Mendes – CCTC, University of Minho (Portugal)
Vítor Costa – University of Porto (Portugal)

Organising Committee

Juan M. Corchado (Chairman) – University of Salamanca (Spain)
Juan F. De Paz (Cochairman) – University of Salamanca (Spain)
Aitor Mata – Universidad de Salamanca (Spain)
Dante I. Tapia – University of Salamanca (Spain)
Javier Bajo – University of Salamanca (Spain)
Sara Rodríguez – University of Salamanca (Spain)
Rosa Cano – University of Salamanca (Spain)
Cristian Pinzón – University of Salamanca (Spain)

Contents

Proteins and Cells

Mathematical Models

Comparing Time Series through Event Clustering*

Juan A. Lara[1], Aurora Pérez[1], Juan P. Valente[1], and África López-Illescas[2]

[1] Facultad de Informática, Universidad Politécnica de Madrid,
Campus de Montegancedo, 28660, Boadilla del Monte, Madrid, Spain
j.lara.torralbo@upm.es, {aurora,jpvalente}@fi.upm.es
[2] Centro Nacional de Medicina del Deporte,
Consejo Superior de Deportes, C/ El Greco s/n, 28040, Madrid, Spain
africa.lopez@csd.mec.es

Abstract. The comparison of two time series and the extraction of subsequences that are common to the two is a complex data mining problem. Many existing techniques, like the Discrete Fourier Transform (DFT), offer solutions for comparing two whole time series. Often, however, the important thing is to analyse certain regions, known as events, rather than the whole times series. This applies to domains like the stock market, seismography or medicine. In this paper, we propose a method for comparing two time series by analysing the events present in the two. The proposed method is applied to time series generated by stabilometric and posturographic systems within a branch of medicine studying balance-related functions in human beings.

Keywords: Data Mining, Time Series, Event, Stabilometry, Posturography.

1 Introduction

Knowledge discovery in databases (KDD) is a non-trivial process that aims to extract useful, implicit and previously unknown knowledge from large volumes of data. Data mining is a discipline that forms part of the KDD process and is related to different fields of computing, like artificial intelligence, databases or software engineering. Data mining techniques can be applied to solve a wide range of problems, including time series analysis, which has come to be highly important in recent years.

A time series can be defined as a sequence X of time-ordered data $X = \{x_t, t = 1,\ldots,N\}$, where t represents time, N is the number of observations made during that time period and x_t is the value measured at time instant t. Time series are usually represented as a graphs in a Cartesian system, where one of the axes represents time and the other (or others in the case of multidimensional series) records the value of the observation (Figure 1).

One interesting problem in the data mining field is the comparison of two time series. This calls for the determination of a measure of similarity indicating how alike two time series are. Most existing techniques compare one whole series with another whole series [1, 2]. However, there are many problems where it is requisite to focus on certain regions of interest, known as events, rather than analysing the whole time

* This work was funded by the Spanish Ministry of Education and Science as part of the 2004-2007 National R&D&I Plan through the *VIIP* Project (DEP2005-00232-C03).

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 1–9, 2009.
springerlink.com

series [3]. This applies, for example, to domains like seismography, where points of interest occur when the time series shows an earthquake, volcanic activity leading up to the earthquake or replications.

In this article, on the one hand, we propose a method that can locate similar events appearing in two different time series, that is, events that are similar and common to the two series, and, on the other hand, we also define a similarity measure between the two time series based on the idea that the more events they have in common the more alike they will be. This similarity measure will be needed to do time series clustering, pattern extraction and outlier detection.

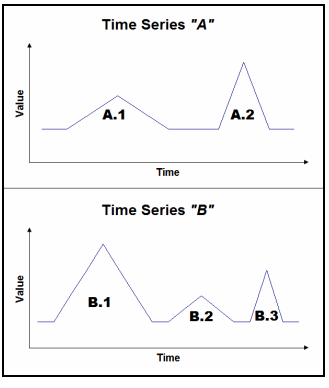


Fig. 1. Charts of two time series (*A* and *B*)

Example. Suppose that the events of a time series in a particular domain are the peaks generated by the local maxima. Given two time series, *A* and *B* (Figure 1), there are two regions of interest in series *A* and three in series *B*. Comparing the two series, we find that the first event in *A* (*A.1*) is very like the second in *B* (*B.2*), and the third event in *B* (*B.3*) is very like the second in *A* (*A.2*). In this case, the series *A* and *B* have two events in common and are, therefore, very alike.

The method developed throughout this article will be applied in the field of medicine known as stabilometry, which is responsible for examining balance-related functions in human beings. This is an important area within neurology, where the diagnosis and treatment of balance-related dysfunctions like *dizziness* has advanced enormously in recent years.

There is a device, called a posturograph, which is used to measure balance-related functions in human beings. The patient stands on a platform to do a number of tests (see Figure 2). We used a static *Balance Master* posturograph. In a static posturograph, the platform on which the patient stands, does not move. The platform has four sensors, one at each of the four corners: front-right (*FR*), front-left (*FL*), rear-right (*RR*) and rear-left (*RL*). While the patient is doing a test, each of the sensors receives a datum every 10 milliseconds. This datum is the intensity of the pressure that the patient is exerting on the above sensor. Therefore, at the end of the test, we have a time series composed of a set number of observations measured at different time points. Each of those observations can be viewed as a point in a four-dimensional space, as the value of each of the four above-mentioned sensors is measured at each time instant.

Fig. 2. Patient doing a test on the posturograph

The proposed method, which will be described in section 3, will be applied to this type of time series. The results of applying the method will be discussed in section 4, whereas the findings will be detailed in section 5. Before all this we will present work related to this article in section 2.

2 Related Work

There are many domains working with time series. Over the last few years, a lot of research has been carried out on the time series field in domains as far apart as medicine [4], the financial sector [5] or traffic control [6].

There are many techniques for comparing time series and extracting common subsequences. A technique for comparing times series based on the Fourier Transform is proposed in [1]. The aim is to extract a number of coefficients from the time series using the discrete Fourier transform, that is, by switching from the time to the frequency domain. Techniques based on alternatives to the Fourier transform, like the Wavelet transform [2, 7, 8], have been proposed.

The landmarks-based technique [9] proposes a method for comparing time series where the singular points (landmarks) of the curve, that is, the maximums, minimums and turning points, have to be stored. It proposes the use of six transformations that are applied to the landmarks. Several features of the landmarks are invariant to the application of certain transformations, and only the invariant features of each transformation are taken into account when looking for series that are similar under certain transformations.

Another type of approach employs the time warping technique. This is based on the idea that two series are similar if the distance between the two series when one of them is compressed on the time axis is less than a certain threshold [10, 11]. Another type of technique uses MBR (minimum bounding rectangles) to compare time series [12].

A method for discovering subsequences common to several series is proposed in [13]. It is based on the idea of building a tree storing all the possible common subsequences of length "*k*" at each depth level "*k*". The technique is very efficient thanks to the tree pruning process, which rejects solutions that do not meet certain minimum confidence conditions.

There are techniques, such as those proposed in [1] and [2], that are useful for comparing whole time series. In many domains, however, as is the case with stabilometry, it is more important to focus on certain events or regions of interest. There are different proposals in this respect, like Povinelli's system [3, 14]. Because of the increasing importance of events detection over the last few years, techniques and algorithms have been developed to detect events in complex time series [15].

In some domains, the value of not one but several observations might be measured at each time point, leading to multidimensional time series. This applies to the posturograph used in this research, as four values are recorded at each time instant. There are several techniques for studying multidimensional time series [16, 17].

In recent years, different and innovative proposals have emerged for data mining time series. Some of these methods use Markovian models to compare time series [18], others use graph theory [19], others again are based on comparing time series by looking at how they change shape [20], etc.

In this paper, we propose the application of computing techniques to medicine, something that has already been done in earlier work. Ever since the early expert systems, like Dendral [21] or Mycin [22], were first conceived, a host of medical decision support software systems have been developed [23, 24].

In [25] a system was developed capable of diagnosing injuries in top-competition athletes thanks to the analysis, using data mining techniques, of data generated by an isokinetics machine that measures the athletes' muscle strength when bending and stretching their members. Similarly, [4] proposes a technique to diagnose epilepsy, using data mining techniques.

The use of computing techniques in the domain of medicine has recently come to be common practice. However, the application of data mining techniques to posturographic data has a number of particularities that, taken together, single it out from their use in other domains. They are: (1) the structural complexity of the patient examinations, (2) the multi-dimensionality of the collected variables and (3) the fact that relevant information appears in definite regions of each series and not across the whole series.

3 Proposed Method

The method presented here is applicable for domains where the points at which a particular event takes place and not the whole time series are of interest. An event is part of the time series that starts at one point and ends at a later point and is of interest for the expert in the domain in question. Many state-of-the-art techniques, like the Discrete Fourier Transform of the Wavelet Transform, compare two time series as a whole. However, these techniques are not suited for those domains where only small parts of the time series are relevant. These events can occur at any instant and with random duration. The method described in this paper focuses on this kind of time series.

Note that the reference domain in this article is stabilometry. The study run focused on one of the tests run on the posturograph: the *UNI* test. This 10-second test aims to measure how well the patient is able to keep his or her balance when standing on one leg with either both eyes open or both eyes shut. While the patient is doing the test, each sensor sends a datum to the central computer every 10 milliseconds. This datum is the intensity or pressure that the patient is putting on that sensor. Therefore, at the end of the test, we have a time series storing four values for each time instant (see Figure 3).

An ideal test would be one where the patient kept a steady stance and did not wobble at all throughout the whole test. The interesting events of this test occur when the patient loses balance and puts the lifted leg down onto the platform. This type of event is known in the domain as a *fall*. When there is a fall, the respective sensors for the lifted leg will register the pressure increase. Figure 3 shows the time series of a patient who has done the UNI test. The curves at the top of the figure are the values recorded by the RR and RF sensors, that is, the right-leg sensors, the leg the patient was standing on. The curves at the bottom of the figure are the values recorded by sensors LR and LF, that is, the left-leg sensors, the leg that should be lifted. The pressures peaks generated when there is a fall event are highlighted.

The next step after identifying the fall events is to determine a similarity measure between two multidimensional time series by obtaining and comparing the fall events that appear in both time series.

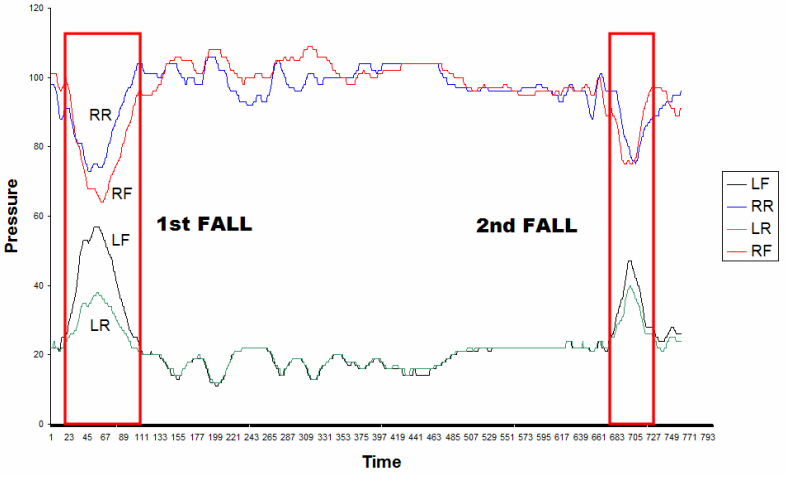


Fig. 3. UNI test time series, highlighting two events (*falls*)

Formally, the aim is to find a function F that takes two times series A and B and returns a similarity value in the interval [0,1], where 0 indicates that the two series are completely different and 1 denotes that the two series are identical, as described in equation (1).

$$F : ST, ST \rightarrow [0,1]$$
$$F : A, B \rightarrow Similarity(A, B) \tag{1}$$

To determine similarity, the proposed method looks for events that appear in both series. The greater the number of events that the two series to be compared have in common, the closer similarity will be to 1. If the series do not have any event in common, similarity will be equal to 0.

To determine whether an event in one time series appears in another, the event has to be characterized by means of a set of attributes and compared with the other events of the other series. To speed up this process, all the events present in the two time series are clustered. Therefore, if two events belong to the same cluster, they are similar. The goal is to find events that are members of the same cluster and belong to different time series.

Therefore, the proposed algorithm for extracting events common to two time series *A* and *B* is:

1. **Extract all the events E_j of both series (events that appear in *A* or in *B*) and characterize each event by means of a set of attributes.** This point is domain dependent, as event characterization will depend on the type of time series. For example, the events of interest in the reference domain we are using are *falls* and they are characterized by the following attributes:
   a) Region in which the lifted leg falls.
   b) Intensity of the pressure exerted by the falling patients' foot on the platform and drop in the intensity of pressure of the standing leg sensors.
   c) Time from when the patient starts to lose balance until he or she falls.
   d) Time from when the patient falls to when he or she recovers.
2. **Cluster all the events extracted in point 1.** To do this, it is necessary to calculate the distance between each pair of events explained under step 1 of the algorithm. We opted to use the city-block distance. This distance calculates the sum of the absolute differences of each of the coordinates of two vectors:

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{2}$$

   In equation (2), *i* and *j* are the vectors to be compared and *p* is the number of coordinates (dimension). We looked at other distance measures, but finally opted to use the city-block distance, because the mean distance per attribute is used during the clustering process to determine whether two elements are members of the same cluster. This mean distance per attribute is obtained straightforwardly by dividing the total distance d_{ij} by the number of attributes *p*.
3. **For each cluster from step 2 and as long as there are events from the two series in the cluster:**
   3.1. **Create all the possible event pairs (E_i, E_k) for which $E_i \in A$ and $E_k \in B$.**
   3.2. **Select the event pair that minimizes *distance* (E_i, E_k).** Equation (2) describes the distance to be used.

(This extracts the two events that are in the same cluster, belong to different time series and are the most alike)

3.3. **Delete events E_i and E_k from the cluster.**

3.4. **Return the pair (E_i,E_k) as an event common to both series.**

By the end of this process, we will have managed to extract the event pairs that are similar in the two series. This is a key point for the mechanism that establishes how alike the two time series being compared are.

A common event, C_i is a pair $C_i = (E_i,E_k) \mid E_i \in A$, $E_k \in B$, output by step 3.4 of the algorithm. If $E = \{E_j, j=1..n\}$ is the set of all events present in A or in B (output by step 1 of the algorithm) and $C=\{C_i, i=1..m\}$ is the set of common events present in both series, A and B, then the similarity can be obtained by comparing the amount of the time series that is common to the two time series (C_i) with the total amount of the time series of interest (E). The more events the series to be compared have in common, the closer the similarity will be to 1.

4 Results

The system implementing the described method has been evaluated by running a battery of tests. These tests were done on time series generated by a posturographic device. The study focused on the data from the *UNI* test. For this test, the events of interest are *falls* that take place at times when the patient suffers a severe loss of balance leading him or her to put down the leg that he or she should be lifting.

We received support for the evaluation from the *High Council for Sports*, an institution attached to the *Ministry of Education and Science*, responsible for coordinating sporting activities in Spain. This institution provided times series for 10 top-competition athletes of different sexes, ages and sports for this study. Note that each time series is composed of four dimensions. At this early stage of the project, 10 is a reasonable number of patients, taking into account how difficult is to obtain this information due to the shortage of top-competition athletes, the complexity of the tests and the fact that there is no public posturographic database. An expert from the above institution helped to validate the results generated by implementing the proposed method. We had to rely on only one expert because stabilometry is a very new field and there are not many experts in this area. In actual fact, there are only a couple of stabilometric devices in use in Spain.

The evaluation of the research focused on one point: *Are the comparisons made by the system of similar quality to those made by the expert?*

To evaluate this point, all time series were compared with each other (if there are 10 time series, and each time series is compared with all the others except itself, we have a total of 45 comparisons). For each of the above comparisons, the similarity rating generated by the method was checked against the similarity score determined by the expert. In each comparison, the expert was asked to determine a similarity rating from the following: {*Not at all similar, Not very similar, Moderately similar, Fairly similar, Very similar*}.

The rating *Not at all similar* would correspond to a similarity in between the interval [0, 0.2), the rating *Not very similar* would correspond to the interval [0.2, 0.4), and

so on up to the rating *Very similar*, which would correspond to a similarity score in [0.8, 1].

When evaluating a comparison, the agreement between the expert and the method could be *Total* if the similarity interval is the same in both cases, *Very High* if the interval determined by the system and by the expert are adjacent, and *Low* in any other case.

The results of the comparisons by the expert and the system were also good, as, agreement between the system and the expert was *Total* or *Very High* in 39 out of 45 cases. Only 6 of the cases showed some differences between the results generated by the system and the ratings determined by the expert.

5 Conclusions

We have developed a method to compare time series by matching up their relevant events. This method is suitable for domains where the relevant information is focused on specific regions of the series, called events, and where the remaining regions are not relevant.

The method was evaluated on time series for top-competition athletes. After performing the different evaluation tests, the results were considered very satisfactory for both the research team and the expert physicians, boosting their will to develop further cooperation in this field.

This project is at a very early stage. The method we have developed is a preliminary version. In the future we intend to refine the method using other distance measures, and apply this method to time series from some other domains.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient Similarity Search In Sequence Databases. In: FODO. Evanston, Illinois (1993)
2. Chan, K., Fu, A.W.: Efficient Time Series Matching by Wavelets. In: ICDE, pp. 126–133. Sydney-AUS (1999)
3. Povinelli, R.: Time Series Data Mining: identifying temporal patterns for characterization and prediction of time series, PhD. Thesis. Milwaukee (1999)
4. Chaovalitwongse, W.A., Fan, Y., Sachdeo, R.C.: On the Time Series K-Nearest Neighbor Classification of Abnormal Brain Activity. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 1 (2007)
5. Lee, C.L., Liu, A., Chen, W.: Pattern Discovery of Fuzzy Time Series for Financial Prediction. IEEE Transactions on Knowledge and Data Engineering 18(5) (2006)
6. Yin, J., Zhou, D., Xie, Q.: A Clustering Algorithm for Time Series Data. In: Proceedings of the Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2006). IEEE, Los Alamitos (2006)
7. Tseng, V.S., Chen, C., Chen, C., Hong, T.: Segmentation of Time Series by the Clustering and Genetic Algorithms. In: Sixth IEEE International Conference on Data Mining - Workshops ICDMW (2006)

8. Kumar, R.P., Nagabhushan, P., Chouakria-Douzal, A.: WaveSim and Adaptive WaveSim Transform for Subsequence Time-Series Clustering. In: 9th International Conference on Information Technology, ICIT (2006)
9. Perng, C., Wang, H., Zhang, S.R., Parker, D.S.: Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases. In: ACDE, San Diego, USA, pp. 33–44 (2000)
10. Rafiei, D., Mendelzon, A.: Similarity-Based Queries for Time Series Data. In: ACM SIGMOD, Tucson, AZ, pp. 13–25 (1997)
11. Park, S., Chu, W., Yoon, J., Hsu, C.: Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases. In: ICDE, San Diego, USA, pp. 23–32 (2000)
12. Lee, S., Chun, S., Kim, D., Lee, J., Chung, C.: Similarity Search for Multidimensional Data Sequences. In: ICDE, San Diego, USA, pp. 599–610 (2000)
13. Alonso, F., Martínez, L., Pérez, A., Santamaría, A., Caraça-Valente, J.P.: Integrating Expert Knowledge and Data Mining for Medical Diagnosis. In: Expert Systems Research Trends, cap. 3, pp. 113–137. Nova Science Ed. (2007)
14. Povinelli, R., Feng, X.: A New Temporal Pattern Identification Method for Characterization and Prediction of Complex Time Series Events. IEEE Transactions on Knowledge and Data Engineering 15(2) (2003)
15. Vilalta, R., Sheng, M.: Predicting rare events in temporal domain. In: IEEE International Conference on Data Mining, pp. 474– 481 (2002)
16. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases, pp. 4190–429. ACM SIGMOD (1994)
17. Kahveci, T., Singh, A., Gürel, A.: Shift and scale invariant search of multi-attribute time sequences, Technical report, UCSB (2001)
18. Wang, Y., Zhou, L., Feng, J., Wang, J., Liu, Z.: Mining Complex Time-Series Data by Learning Markovian Models. In: Proceedings of the Sixth International Conference on Data Mining ICDM 2006. IEEE, Los Alamitos (2006)
19. Jan, L., Vasileios, L., Qiang, M., Lakaemper, W.R., Ratanamahatana, C.A., Keogh, E.: Partial Elastic Matching of Time Series. In: Proceedings of the Fifth IEEE International Conference on Data Mining (2005)
20. Dong, X., Gu, C., Wang, Z.: Research On Shape-Based Time Series Similarity Measure. In: Proceedings of the IEEE Fifth International Conference on Machine Learning and Cybernetics. Dalian (2006)
21. Lederberg, J.: How Dendral Was Conceived and Born. In: ACM Symposium on the History of Medical Informatics, November 05 1987. National Library of Medicine, Rockefeller University (1987)
22. Shortliffe, E.H.: Computer Based Medical Consultations: MYCIN. American Elsevier, Amsterdam (1976)
23. Edberg, S.C.: Global infectious diseases and epidemiology network (GIDEON): A world wide web-based program for diagnosis and informatics in infectious diseases, Clinical Infectious Diseases. official Publication of the Infectious Diseases Society of America 40(1), 123–126 (2005)
24. Gil, D., Soriano, A., Ruiz, D., Montejo, C.A.: Embedded systems for diagnosing dysfunctions in the lower urinary tract. In: Proceedings of the 22nd Annual ACM Symposium on Applied Computing (2007)
25. Alonso, F., Caraça-Valente, J.P., González, A.L., Montes, C.: Combining expert knowledge and data mining in a medical domain. Expert Systems with Applications 23, 367–375 (2002)

Visual Knowledge-Based Metaphors to Support the Analysis of Polysomnographic Recordings

Abraham Otero[1], Paulo Félix[2], and Carlos Zamarrón[3]

[1] Department of Information and Communications Systems Engineering, University San Pablo CEU, 28668 Madrid, Spain
abraham@dec.usc.es
[2] Department of Electronics and Computer Science, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
paulo@dec.usc.es
[3] Division of Respiratory Medicine, University Hospital Complex of Santiago de Compostela, 15752 Santiago de Compostela, Spain
carlos.zamarron.sanz@sergas.es

Summary. This paper presents algorithms that provide support in the task of reviewing the physiological parameters recorded during a polysomnography, the gold standard test for the diagnosis of Sleep Apnea-Hypopnea Syndrome (SAHS). This support is obtained through the generation of visual metaphors which help identify events (apneas, hypopneas and desaturations) that occur over the span of the recording and are relevant to the diagnosis of SAHS.

The definition of these events is not completely standardized and it is not unusual that different physicians use different criteria when identifying them. To tackle this problem our algorithms start with a linguistic description of the events to be identified. This description is obtained directly from the clinical staff and is projected onto a set of algorithms of a structural nature that support the generation of the visual metaphors. To represent and manipulate the imprecision and vagueness characteristic of medical knowledge we rely on the fuzzy set theory.

The metaphors proposed herein have been implemented in a tool aimed at supporting the diagnosis of SAHS. The tool provides wizards that permit the morphological criteria that define the apneas, hypopneas and desaturations to be customized by the physician and the visual metaphors automatically reflect the new criteria.

1 Introduction

Sleep Apnea-Hypopnea Syndrome (SAHS) is a common sleep-breathing disorder characterized by recurrent episodes of the upper airway narrowing or collapsing during sleep. An obstruction is caused by the soft palate and/or base of the tongue collapsing against the pharyngeal walls. When the obstructions are complete they are called apneas; when they are partial, hypopneas. It is estimated that SAHS affects 4% of the adult male population and 2% of the adult female population [15], having an especially high prevalence in adult males with obesity problems, and it is recognized as an important public health issue [12].

Apneas and hypopneas are accompanied by hypoxemia, with a drop in SpO2, surges in blood pressure and brief arousal from sleep. Arousals do not necessarily

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 10–20, 2009.
springerlink.com

take the patient to a conscious state, but they make him/her leave the deeper sleep stages – i.e., stages where sleep has more refreshing effects – and make him/her spend a higher fraction of nightly rest in stages closer to vigil. As a result, the patient's sleep architecture is fragmented and its refreshing effects are diminished. Thus, patients often suffer diurnal somnolence and cognitive deficits that increase the risk of working and driving accidents [3]. They may also suffer from depression, anxiety, excessive irritability and several sexual dysfunctions.

Overnight polysomnography is currently considered the diagnostic gold-standard for SAHS. It is performed in a hospital Sleep Unit and consists of the registration of a wide range of physiological parameters while the patient is asleep: respiratory airflow (RA), blood oxyhemoglobin saturation (SpO2), respiratory effort, electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), etc. The diagnosis of SAHS is a tedious undertaking that requires visual inspection, usually with the assistance of a computer, of the long signal recordings obtained during the polysomnography.

Our goal is to develop a set of visual metaphors to provide support in the task of inspecting a polysomnographic recording. To this end, the metaphors try to highlight the most relevant events in the diagnosis of SAHS: apneas, hypopneas and desaturations. One of the challenges of this task is the lack of a universally accepted definition of such events: different physicians may identify them using different criteria. This had led us to start with the linguistic definition of the events that the physician feels more comfortable with, and to adapt the visual metaphors accordingly.

Section 2 describes the representation we will use for time and it summarizes some basic fuzzy concepts on which our algorithms are based. Sections 3 and 4 present the algorithms which provide support for the generation of the visual metaphors aimed at simplifying the identification of apneas and hypopneas, and desaturations, respectively. These algorithms are of a structural nature and they take advantage of the fuzzy set theory to model and represent medical knowledge close to human intuition. A desktop tool that implements the aforementioned techniques is presented in section 5. Finally, the results obtained are discussed and a series of conclusions are given in sections 6 and 7, respectively.

2 Prior Definitions

We consider time as being projected onto a one-dimensional discrete axis $\tau = \{t_0, t_1, ..., t_i, ...\}$ such that for every $i \in \mathbb{N}$, $t_{i+1} - t_i = \Delta t$, where Δt is a constant. Δt is the minimum step of the temporal axis. Thus given an i belonging to the set of natural numbers $\mathbb{N}$, t_i represents a *precise* instant.

Given as discourse universe the set of real numbers $\mathbb{R}$, a **fuzzy number** A is a normal and convex fuzzy subset of $\mathbb{R}$ [5]. A fuzzy set A with membership function μ_A is *normal* if and only if $\exists\ v \in \mathbb{R}$, $\mu_A(v) = 1$. A is said to be *convex* if and only if $\forall v, v', v'' \in \mathbb{R}$, $v' \in [v, v'']$, $\mu_A(v') \geq \min\ \mu_A(v), \mu_A(v'')\}$.

Normality and convexity properties are satisfied by representing π_A, for example, by means of a trapezoidal representation. In this way, $A = (\alpha, \beta, \gamma, \delta)$,

$\alpha \leq \beta \leq \gamma \leq \delta$, where $[\beta, \gamma]$ represents the core, $core(A) = \{v \in \mathbb{R} |\ \pi_A(v) = 1\}$, and $]\alpha, \delta[$ represents the support, $supp(A) = \{v \in \mathbb{R} | \pi_A(v) > 0\}$ (see Fig. 1). We have opted for this representation for possibility distributions on the basis of its computational efficiency and the intuitiveness of its semantics for medical users.

We obtain a fuzzy number A from a flexible constraint given by a possibility distribution π_A, which defines a mapping from $\mathbb{R}$, to the real interval $[0, 1]$. A fuzzy constraint can be induced from a piece of information such as emph"x has a high value", and given a precise number $v \in \mathbb{R}$, $\pi_{A=high}(v) \in [0, 1]$ represents the possibility of x being precisely v. By means of π_A we define a fuzzy subset A of $\mathbb{R}$, which contains the possible values of A, being A a disjoint subset.

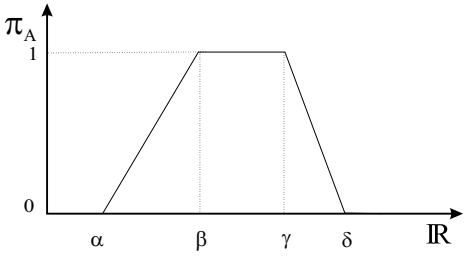


Fig. 1. Trapezoidal possibility distribution

3 Providing Support for the Identification of Apneas and Hypopneas

In the bibliography there is no absolute consensus on the criteria to be fulfilled by an episode of apnea [13]. The criteria which is probably most widely accepted in Europe is a reduction in the volume of air inhaled to at least 10% of the basal value sustained for at least 10 seconds. However, some authors defend that the reduction in the airflow must be to at least 5% [6], while the American Academy of Sleep Medicine (AASM) requires a total cessation [9]. The latter criteria leaves the problem of noise: it is not uncommon that even when there is no airflow the RA signal presents a non null value because of the presence of noise.

Hypopneas are associated with a reduction in RA to at least 50% of the basal value. For some physicians this event is enough to label the hypoventilation as an episode hypopnea, while others require the hypoventilation to provoke an arousal and/or a drop in SpO2. Under certain conditions, the AASM also labels reductions in RA of just 70% of the basal value as hypopneas.

The algorithms we have developed do not commit to any criteria. Our goal is to project the event's linguistic description with which the physician feels more comfortable onto a computational representation. Then this representation will be used to create a visual metaphor to assist in the task of identifying those fragments of the RA in which there has been a reduction of flow compatible with the description. This visual metaphor will take the form of a semitransparent grid drawn over the RA.

3.1 The Algorithm

We start by filtering the RA signal with a third-order Butterworth bandpass filter with cut-off frequencies of 0.20 Hz (one breath every 5 seconds) and 0.45 Hz (one breath every 2.2 seconds). On average, during the night the patient breathes every three seconds. In order to obtain a null-phase filter, after filtering in the forward direction, the filtered sequence is then reversed and run back through the filter; the filtered signal is the reverse of the output from the second filtering operation.

The RA signal presents continuous oscillations corresponding to inhalations and exhalations of the patient (see Fig 2). Its instantaneous value is not relevant to the study of SAHS; its envelope is what really reflects the air that is being inhaled or exhaled. Thus we start by obtaining a signal, which we shall call V, whose value at every instant attempts to reflect the amount of air that is being inhaled or exhaled. To this end we search for the maximum and minimum of RA within a mobile window of 1.5 seconds (the approximate length of an inhalation or exhalation), and calculate the difference between the two values. This difference will be the value of V during the 1.5 seconds window. The value of this signal is directly proportional to the instantaneous amount of air being inhaled or exhaled.

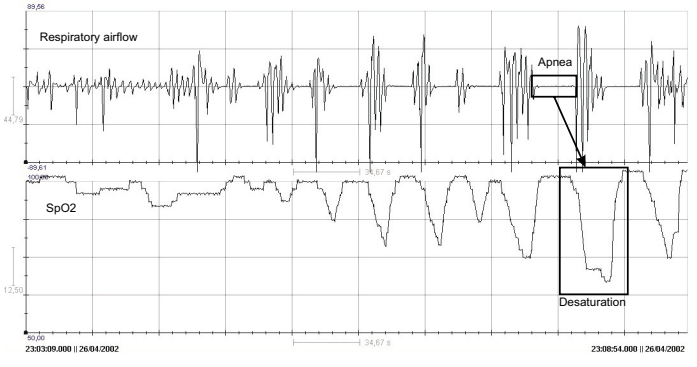


Fig. 2. Fragment of a polysomnogram with several apneas and the corresponding desaturations. The association between one of the apneas and its desaturation is shown.

The calculation of the basal value is challenging: both the different sleep stages and the patient postural changes during the night affect the amplitude of RA. Therefore, it is not acceptable to calculate a basal value considering the whole recording, or its first few minutes. The use of a mobile window does not provide good results either. In the basal value calculation only normal breathing should be considered, while those intervals containing hypoventilations should be ignored. If not, the calculated value will be less than the one corresponding with normal breathing.

To overcome this problem for each sample $RA[t_i]$ we take a two minute window centered on it, in order to calculate the basal value. For each sample of the

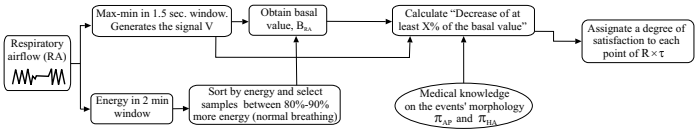


Fig. 3. Block diagram of the algorithm which supports the visual metaphor for highlighting apneas and hypopneas

previous window we calculate the instantaneous energy of RA using a 3 seconds window centered over the sample itself. Then the energy values obtained are sorted from lowest to highest. The time instants corresponding to the samples whose energy falls between 80% and 90% higher energy are selected. Finally the basal value for each sample, $B_{RA}[t_i]$, is estimated as the average value of the samples of V corresponding with the selected time instants (see Fig. 3).

By ignoring those samples corresponding with energy values under 80% we are trying to discard the apneas and hypopneas in the basal value calculation. Given that in the intervals where these hypoventilations have occurred there is a reduction in the amplitude of the signal RA, the energy of the signal in such intervals is expected to be low. By ignoring those samples with energy values above 90% we are trying to discard possible artifacts that the signal may contain.

Fuzzy logic has proven its capabilities to represent and manipulate the vagueness characteristic of medical knowledge [2]. We shall take advantage of this formalism to represent the linguistic descriptions of apneas and hypopneas. More specifically, we shall represent the percentages of reduction from the basal value that the physician wants to use for the identification of these events by means of the trapezoidal possibility distributions D_{AP}, for apneas, and D_{HA}, for hypopneas. For example, if a physician defines apnea as a "*decrease to at least approximately 10% from the basal value*", the percentage of reduction "*decrease to at least approximately 10%*" can be represented by the trapezoidal possibility distribution $D_{AP} = (0, 0, 0.1, 0.15)$. Thus "*decrease to at least approximately 10% of its basal value*" can be obtained as: $D_{AP} \otimes B_{RA}[t_i]$, where $\otimes$ represents the fuzzy product of the fuzzy value D_{AP} by the basal value of RA in t_i, $B_{RA}[t_i]$ [5].

The RA signal presents continuous oscillations which are approximately symmetrical around the x-axis. Given a point $(y, t_i) \in \mathbb{R} \times \tau$, the possibility of the patient experiencing an apnea if y is the maximum or the minimum value of the RA oscillation which contains t_i will be given by:

$$\pi_{AP}(y, t_i) = \mu_{D_{AP} \otimes B_{RA}[t_i]}(|y|) \tag{1}$$

where $|y|$ is the absolute value of y. This expression allows us to obtain the degree of compatibility of each point of $\mathbb{R} \times \tau$ with the linguistic expression represented by D_{AP}, i.e., with a reduction in airflow compatible with an apnea (see Fig. 3). To obtain the degree of compatibility of a point of $\mathbb{R} \times \tau$ with the criteria used for a reduction in airflow compatible with hypopnea we use a similar expression:

$$\pi_{HA}(y, t_i) = \mu_{D_{HA} \otimes B_{RA}[t_i]}(|y|) \tag{2}$$

where D_{HP} represents the percentage of a reduction in RA compatible with hypopnea: D_{HP} = "*decrease to at least approximately 50%*".

3.2 The Visual Metaphor

Our goal is to build a desktop application which provides support for the reviewing of polysomnograms. This tool must plot the polysomnogram signals; each of them is represented in a rectangular area of the screen which we shall call channel. When the channel represents RA, the 0 magnitude value will be placed in the middle of the channel, so that the screen pixels of the upper half of the channel correspond to positive values and the pixels in the bottom half correspond to negative values.

There is a one to one correspondence with each pixel of the channel and a point of $\mathbb{R} \times \tau$. Therefore, using equations 1 and 2 we can calculate the degree of compatibility of each pixel of the RA channel with the apnea and hypopnea criteria. A graphical representation appropriate for this information can be a color code that represents the compatibility of each pixel with each of the two criteria. In our tool the color red has been associated with the total compatibility of the apnea criteria and yellow with the null compatibility. To obtain colors corresponding to intermediate levels of compatibility the red color is degraded to yellow using a linear gradient, and a mapping between colors and compatibilities is generated. The total compatibility for the hypopnea criteria has been associated with yellow, and this color is degraded to green, when compatibility for this criteria becomes null, using a linear gradient. Our tool makes a semitransparent drawing of this information over the RA channel. To this end we have taken advantage of the capabilities of the Java2D graphics library.

The information about the compatibility with the apnea and hypopnea criteria can be represented in two different copies of the RA signal. However, no matter which definition is used, the region where the apnea criteria presents non null compatibility will always be located inside the region of maximum compatibility with the hypopnea criteria. Thus we have chosen to superimpose both representations. A gray line is drawn where the compatibility of the apnea criteria becomes null, so the physician can easily identify this point.

Fig. 4 shows a fragment of a polysomnogram where the RA signal is displayed using the grid described herein. Approximately in the middle of the fragment, the amplitude of the RA oscillations changes. Note how at the beginning both the regions corresponding with apneas and with hypopneas are wider, and they tighten towards the end, adapting themselves to the new basal value of RA.

4 Providing Support for the Identification of Desaturations

There is no absolute consensus on the criteria to be fulfilled by a relevant drop in SpO2, although this time the differences are more subtle: the criteria usually requires a fall in SpO2, relative to the basal value of the signal, of 3%, 4% or

Fig. 4. Tool which implements the algorithms and permits the modification of the morphological criteria that define the relevant events in the SAHS study

5%. Again, our algorithms do not commit to any criteria but seek to capture that one with which the physician feels more comfortable.

4.1 The Algorithm

SpO2 basal value can also vary throughout the night. Long periods of hypoventilation may produce a decrease in the basal value, which can recover if the patient breathes normally for a sufficient amount of time. For the calculation of the SpO2 basal value our algorithms use a mobile window of 10 minutes centered over each sample of the signal. The values of SpO2 in this window are sorted from highest to lowest, and the basal value for the sample SpO2$[t_i]$, $B_S[t_i]$, is obtained as the average of the values of the 10% samples with higher value inside the window.

The artifacts that occur over the SpO2 always produce null values. By considering only the 10% samples with higher value we are ignoring the episodes of desaturation -whose samples will fall within the 90% smaller values-, and the possible artifacts.

Given a point $(y, t_i) \in \mathbb{R}^+ \times \tau$ -SpO2 is always positive-, the possibility of the patient experimenting a desaturation in t_i if SpO2$[t_i] = y$ will be given by:

$$\pi_{Des}(y, t_i) = \min\{\mu_C(B_S[t_i] - y)\} \tag{3}$$

where C is a trapezoidal possibility distribution that represents the value of a drop in SpO2 compatible with a desaturation. For example, if the linguistic criteria that the physician prefers is "*fall of more than approximately 4%*" then C can be represented by $(3, 4, 100, 100)\%$.

4.2 The Visual Metaphor

Eq 3 can be used to calculate the degree of compatibility of each screen pixel of the SpO2 channel, which corresponds to a point in $\mathbb{R}^+ \times \tau$, with the criteria chosen by the physician. This compatibility will be represented, again, by a color code where red represents the maximum compatibility and green, the minimum. A linear gradient is used to obtain a set of intermediate colors between red and green corresponding with compatibilities between 1 and 0.

Although it is not a standardized polysomnographic criteria, the team of pneumologists which works with us prefer to use $C = (3, 20, 100, 100)\%$, whose linguistic meaning is "*drop in SpO2 of moderately high severity*". With this criteria our visual metaphors will assign a low possibility value to small drops in SpO2 compatible with a desaturation, but they still will be highlighted. Drops of high severity (drops of more than 20 %) are assigned the total compatibility. Thus, the metaphor provides visual feedback not only on the number and position of the desaturations, but also on their severity.

In Fig. 4 we can see how the patient's SpO2 basal value presents a decrease approximately in the middle of the recording, probably as consequence of the decrease in amplitude of the oscillations of RA. Note how the grid also falls in the middle of the recording, adapting itself to the new SpO2 basal value.

5 Experimental Results

All algorithms presented in this paper, as well as the visual metaphors they support, have been implemented in a desktop tool (see Fig. 4). This tool is capable of loading polysomnographic recordings that are stored in MIT-BIH format. The parameters of the algorithms that represent the different morphological criteria that apneas, hypopneas and desaturations must fulfill can be customized by means of visual wizards (see Fig. 5), and the effects of the customization are immediately reflected on the visual metaphors.

The tool has been tested by physicians belonging to the Division of Respiratory Medicine of the University Hospital Complex of Santiago de Compostela. They have found that the visual metaphors that we have developed provide an

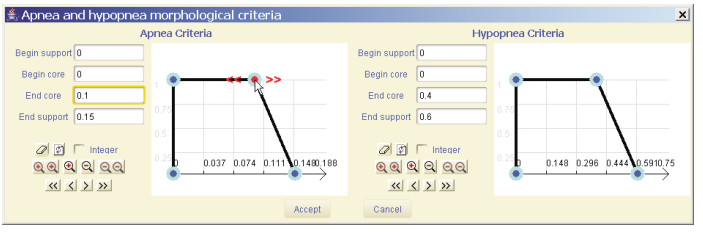


Fig. 5. Wizard which allows the customization of the morphological criteria that define apneas and hypopneas. The trapezoidal possibility distributions can be modified with the mouse or by typing the values in the text fields.

effective support in the identification of episodes of apneas, hypopneas and desaturations. They also consider the metaphors useful in the task of estimating the duration of the events and their severity (percentage of reduction of RA or magnitude of the drop in SpO2). Thus, the physicians are optimistic regarding the potential of the techniques presented in this paper as a support for the diagnosis of SAHS.

A video of the tool showing a polysomnographic recording with the visual metaphors proposed here can be found in [10] . We recommend that the reader watch that video, since it is more effective at demonstrating the visual metaphors we have developed than printed media.

6 Discussion

In the bibliography there are several works which propose a diagnostic test to determine automatically whether or not a patient suffers from SAHS [11]. These test usually take one or several of the physiological parameters recorded during a polysomnography as input. There also are works which propose techniques to identify each episode of apnea and hypopnea individually [1, 4, 7, 8, 14], as well as several commercial solutions with similar capabilities. From this data it can be calculated the apnea-hypopnea index (AHI), which is defined as the number of apneas and hypopneas that the patient suffers per hour of sleep. Based on this index there are medical criteria for associating different levels of severity to the patient's condition.

While these tools provide a significant support for the physician, it is still too early to consider eliminating physicians' intervention in the diagnosis of SAHS. On the one hand, the medical community has been reluctant, probably with good judgment, to trust the results given by automated analysis tools without a contrast. On the other hand, these tools are not able to generate all the information that a physician takes into consideration when diagnosing a patient.

The severity of the condition of two patients who have the same AHI can vary substantially if the average duration of apneas and hypopneas, the average percentage reduction of respiratory airflow during them and the severity of the desaturations are significantly different. When the physician inspects a polysomnographic recording he/she performs a characterization of the various pathological events suffered by the patient. This characterization plays a major role in the final diagnosis. Usually, the tools that perform an automated analysis of the recordings do not provide any of this information, and when they do it is incomplete and unreliable. Thus, at present polysomnographic recordings are always inspected visually by a physician before a diagnosis is issued and a therapy is suggested. Therein lies our interest in providing tools which simplify the reviewing process by helping to identify events relevant in the SAHS diagnosis.

The visual metaphors that we have created also help in the task of characterizing apneas, hypopneas and desaturations. In the first two events, the grid we have developed serves as an aid to measure the duration of the events, as it helps

to identify their beginning and their end. It also helps measure the percentage of reduction from the basal level, as it delimits the regions of the screen which correspond to reductions in RA compatible with apneas and with reductions compatible with hypopneas.

In the case of the SpO2 grid, it helps in measuring the duration of the desaturation, and provides visual feedback on the severity of the drop. The team of physicians which tested the tool decided to use the criteria "*drop in SpO2 of moderately high severity*"to describe the desaturations, instead of a standard one. This reflects their interest in having tools which assist them not only in identifying the events, but also in characterizing them.

7 Conclusions

We have presented a set of structural algorithms whose purpose is to provide support in the reviewing of polysomnographic recordings. The algorithms try to facilitate the identification of apneas, hypopneas and desaturations by generating visual metaphors that help locate these events over the span of a recording. To this end we project a linguistic description of the morphology of these events onto a computational representation. The fuzzy set theory has provided invaluable support in this task. The computational representation is used to calculate the compatibility of each screen pixel of the channel where the signals RA and SpO2 are displayed with the events' linguistic description. This information is represented by a semitransparent grid where different colors represent different levels of compatibility with the criteria.

Our proposal has been implemented in a desktop tool which allows the clinical staff to edit the morphological criteria of the events to be highlighted. Thus, the tool provides support for the use of customized criteria in the analysis of polysomnographic recordings. This solves the problem of the lack of a universal agreement on these criteria in the bibliography, by allowing each physician to use the criteria he/she feels more comfortable with.

For future work, we intend to create new metaphors to identify other events that are recorded during a polysomnography and which, despite having a lower relative importance than the events covered in this paper, are also considered by physicians in the diagnosis of SAHS. Among these events are muscular activity recorded in the electromyography, pauses in respiratory effort and snoring.

Acknowledgments

The authors wish to acknowledge the support by the Spanish Ministry of Education and Science (MEC), the European Regional Development Fund of the European Commission (FEDER) and University San Pablo CEU through grants TIN2006-15460-C04-02 and USP 04/07, respectively.

References

1. Al-Ani, T., Hamam, Y., Fodil, R., Lofaso, F., Isabey, D.: Using hidden Markov models for sleep disordered breathing identification. Simulation modelling 12, 117–128 (2004)
2. Barro, S., Marín, R., Palacios, F., Ruíz, R.: Fuzzy logic in a patient supervision systems. Artificial Intelligence in Medicine 21, 193–199 (2001)
3. Beebe, D.W., Gozal, D.: Obstructive sleep apnea and the prefrontal cortex: towards a comprehensive model linking nocturnal upper airway obstruction to daytime cognitive and behavioral deficits. Journal of Sleep Research 11(1), 1–16 (2002)
4. Cabrero-Canosa, M., Castro-Pereiro, M., Graña Ramos, M., Hernández-Pereira, E., Moret-Bonillo, V., Martin-Egaña, M., Verea, H.: An intelligent system for the detection and interpretation of sleep apneas. Expert Systems with Applications 24, 335–349 (2003)
5. Kaufmann, A., Gupta, M.M.: Introduction to Fuzzy Arithmetic. Van Nostrand Reinhold Company Inc. (1984)
6. Köves, P.: Obstructive Sleep Apnea Syndrome. Springer, Heidelberg (1999)
7. Morsy, A.A., Al-Ashmouny, K.M.: Sleep apnea detection using an adaptive fuzzy logic based screening system. In: 27th IEEE EMB Conference, pp. 6124–6127 (2005)
8. Nazeran, H., Almas, A., Behbehani, K., Lucas, E.: A fuzzy inference system for detection of obstructive sleep apnea. In: 23th IEEE EMB Conference, pp. 1645–1648 (2001)
9. American Academy of Sleep Medicine Task Force. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. Sleep 22, 667–689 (1999)
10. Otero, A.: Visual knowledge-based metaphors to support the analysis of polysomnographic recordings, Video (2008), http://biolab.uspceu.com/sahs
11. Penzel, T., McNames, J., de Chazal, R., Raymond, B., Murray, A., Moody, G.: Systematic comparison of different algorithms for apnea detection based on electrocardiogram recordings. Med. Biol. Eng. Comput 40, 402–407 (2002)
12. Phillipson, E.A.: Sleep apnea. A major public health problem. N. Engl. J. Med. 328, 1271–1273 (1993)
13. Ulukavak, T., Kokturk, O., Ozkan, S.: Apnea-Hypopnea Indexes Calculated Using Different Hypopnea Definitions and Their Relation to Major Symptoms. Sleep and Breathing 8(3), 141–146 (2004)
14. Varaday, P., Micsik, T., Benedeck, S., Benyo, Z.: A nobel method for the detection of apnea and hypopnea events in respiration signals. IEEE Transactions on Biomedical Engineering 49, 936–942 (2002)
15. Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., Badr, S.: The occurrence of sleep-disordered breathing among middle-aged adults. Med. Biol. Eng. Comput. 328(17), 1230–1235 (1993)

A Bio-inspired Proposal for Focus Attention While Preserving Information

O. Bolívar Toledo, J.C. Quevedo Losada, and J.A. Muñoz Blanco

Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
Campus de Tafira s/n, 35017, Las Palmas de Gran Canaria, Spain
obolivar@dis.ulpgc.es, jquevedo@dis.ulpgc.es,
jamunoz@dis.ulpgc.es

Abstract. The interplay between the design of layered computing systems and the study of real visual structures and functions is a very fruitful area in which the characteristics of the natural systems inspire properties and design criteria of the artificial counterparts. Visual saliency and attention mechanism are two of the most explored and interesting features of the visual processing in higher vertebrates. In general, visual saliency refers at the propensity of some visual features to attract attention, while attention by itself requires a voluntary effort to locate places of interest in the visual scene. In this paper we present a study of a mechanism of attention that has variable selection criteria based in highlighting some feature or object in the image while blurring the rest of it. The fact that all the information is preserved in the process, allow us to be able to select a new location of interest in a efficient way. This mechanism is based in a class of transform which combines partitions of the data field and functionals acting on them and is suitable to perform a selection process of the attended location with different levels of resolution.

Keywords: visual attention, visual information, retinal mechanism, computational modelling.

1 Introduction: The Problem of Visual Attention

The study of visual systems from the point of view neurophysiological and from the computational one has always had a twofold purpose namely: to unravel the mysteries of the visual operation in terms of visual encoding and transmitting information, and also to apply this knowledge in the construction of artificial systems that mimic that operational performance and enable its application in robotic systems.

Several neurophysiological and psychological experimental data refer to the hugeness of our perceptual system capacity to process visual information. For many authors it is virtually without limits. The same data also confirm the possibility of a high degree parallel process. From many years, some research has tackled the computational problem of the visual representations in terms of Information Theory concepts such as redundancies reduction, efficient coding, etc [1]; this approach justifies the extended use of orthogonal transforms in visual recognition [2]. However, the investigations oriented in the above sense does not consider what and how the eye sees, not even paying attention to the information selection mechanism nor to the inhibition ones of the irrelevant information. In this sense one of the most explored and interesting feature of the visual processing in

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 21–29, 2009.
springerlink.com

higher vertebrates is the selection process that underlies the decision of which of the information imputing through our eyes receives further processing. This fact plays a central role in at least three basic animal behaviours, which are: alerting, orienting and executive control [3]. The plural nature of what we understand by visual attention, has been the cause of the formulation of multiple descriptions about the attentional mechanisms from different points of view, but however, in spite all the effort carried out from different explaining levels, the initial controversy outlined by James in 1890, [4] about the double nature of attention as a causal mechanism (an agent) or as an effect determined by selective process, has not been yet solve satisfactorily [5].

From the computational point of view, most of the experimental results have confirmed the tendency to consider a framework with two attentional strategies: a first strategy bottom-up, based in the own visual scene [5] [6], which provides with a topographical map of salient clues, followed by a second top-down strategy, based in higher levels mechanisms which provide with depending tasks cues. In this way, certain features of the visual environment becomes in a more or less automatically way in an "attention atractor", giving raise to an outstandingly location or saliency, while others locations would require a volitive effort in being attended. What it seems to be beyond any doubt is that the traditional dichotomy automation versus control has to be conceived as a continuum and the process denominated automatics are also subject in any way to a certain control form. In this context, the phenomenon of visual saliency refers to the propensity of some visual features to attract attention. Visual saliency is determined from a variety of factors. At the lowest levels, color opponencies, unique orientations and luminance contrasts create the effect of visual pop-out. At the highest level it has been proposed that we can prime our visual processes to help guide what we wish to search for in a visual scene. It seems that at this level the process is more conditioned by the knowledge structure that by the properties of picture data.

2 Computational Aspects of the Visual Attention Process

As earlier as the 40s, Norbert Wiener suggested that the "no-man's land" between large bodies of knowledge were the most fertile areas in which the technological and scientific advances would take place in the future. It is in this "no man's land" in which our research is rooted, with the twofold purpose mentioned earlier in section 1: trying to uncover the secrets promised by the nervous system at the visual attention level and the design of artificial counterparts of the observed processes. Hence our interest in finding paradigms or constructs¸ from the computational point of view, applies to these two aspects. We focus our attention in mechanisms such as: computing layers, presynaptic inhibition schemes, analysis of the convergence and divergence of information and the theory of data fields transformations.

2.1 Layered Computation

The concepts of layers computation have been widely studied and used by numerous researchers [7] and its adjustment for the modelling of the Nervous System has given rise to important discoveries [7] [8]. At the moment we can affirm that the nervous

system, as much of the sensorial and cortical subsystems, process the visual information by layers, through complexes and half-framed longitudinal and cross-sectional mechanisms, in which as we move from the sensors towards the cortical zones increases the degree of semantic complexity, making necessary an analysis as much of the neuronal cooperation as of their morphogenesis.

A concept strongly related to the one of computation or process by layers is that of parallel computation, in which different parallel areas have independent functions and is this duality structure-function the one that has given rise to that the operational Systems of Receptive Fields and the transformations on them constitutes one of the more important computational mechanisms in the process of visual data. From photoreceptors to ganglion cells, visual attention is properly interpreted by means of linear and nonlinear spatio-temporal filters center-periphery receptive fields and analogic computation [9] [10] [11].

In this proposal, data processing is performed by layers of similar computational elements. These layers may correspond to functional layers in the retina or might be of many computational layers to one functional retinal layer. Layers may perform either linear or non-linear computation. Linear computation is of the lateral interaction type, being characterized by a weighting function in space and time. Our computational model of the pre-attentive stage assumes that the neuron performs a spatial integration over its receptive field, and that its output activity is a possibly nonlinear function of

$$\iint_{RF} K(x,y)I(x,y)dxdy$$

where I(x, y) is the input, and K(x, y) is a weighting kernel that describes the relative contribution of different locations within the receptive field to the output.

Many of the aspects of visual attention that were previously mentioned are present in this computational construct by layers, because almost all the characteristics that until now seems to be more influential in visual saliency, namely: color opponencies, unique orientations and luminance contrast, take place at the retinal and lateral geniculate nucleus level (in vertebrate visual systems) and may be modelling by selecting an appropriate impulse response. At higher levels (visual cortex) although it is still accepted the structure by layers, a suitable formulation of the process of visual attention requires a nonlinear formulation in which volitive aspects must be modeled at the algorithmic level, by means of cooperative processes, heuristic methods and possibly, at this end, the usual analytic tools of the low level would not be adequate.

2.2 Presynaptic Inhibition Schemes

There are enough evidences that demonstrate that attention process or capacity, besides enhancing the relevant information, also inhibes the potentially distractor one. Differents works related with the attentional focus describe a perifocal zone affected by inhibitory process, where the attentional acuity is worse that in zones remotely situated with respect to the center of attention. [12] .

The presynaptic inhibition mechanisms have been very important to understand the retinal coding, both at the IPL level (inner plexiform layer) [11] and also to explain various visual phenomena, such as the possibility of obtaining invariant representations against global illumination changes in the image impacting on the retina

[13]. They may also help us to explain certain phenomena inherent in the process of visual attention, whose aim would be to highlight a feature or object in the image at the expense of the rest, losing resolution, assuming it does not contain anything relevant. Presynaptic inhibition has an essential role as a selector of one set of stimuli that result in higher areas of the brain, which is equivalent to a high level information filter, in the sense that it filters that semantic content of the information that reached advanced stages of process [11].

2.3 Convergence and Divergence of Information Process

An inherent characteristic of the nervous system is the exhibition of a remarkable convergence and divergence of the traffic signal. The process of convergence and divergence of information have been extensively studied and have interesting properties in the context of reliability. Convergence is necessary if one is to be able to compute over a parallel input domain, and divergence is necessary to recover resolution. In each layer it is observed a high degree of overlap between neighbouring receptive fields. This overlap of receptive fields and the convergence and divergence of information, has inspire the formulation of various models of location and detection of stimuli [7] [8], preservation of information and their relationship to completeness of transformations [14], perception of movement and functioning of ganglion in amphibian retinas [8].

In previous work we have shown how visual information can be processed and preserved in detail through overlapping receptive fields [1,7,8] that vary from layer to layer, presenting a continuous expansion and contraction of the places where that information is processed.

2.4 Complete Transforms

Completeness is a basic concept in the theory of information and it refers to the fact that there is no data loss in the process of signal codification and transformation, or that any loss was offset by an increase in the semantic content of the signal. A study on the transformation process in image and retinal process led us to redefine several concepts related with completeness and with the structure and function of a transformation. This has led to the definition of Resolution Progressive Transformations (TRP), whose immediate property consists in obtaining a space output, which represents the entry space with several resolutions. This equates to dispose at the transformed domain with a sub-images set of different resolution, equivalent to a sampling variable [15]. Based on the study of these transformations [16] we have developed a series of theorems, lemmas and demonstrations related with complete transformations that combine different configurations of receptive fields with functional acting on then. From the analytical standpoint, the complete description requires a priori conservation in the number of degrees of freedom. That constancy has allowed us to establish a sort of principle of conservation which is determined by the functional computed and by the receptive fields, which also reflect a dual situation, in the sense that completeness requires increasing the number of them if the other decreases and vice versa.

3 The Proposed Model: Focus Attention While Preserving Information

As we have mentioned previously, several presently and successfully attentional control bottom-top models are based on a saliency map, being the strategy used to reduce the amount of input sensorial data, and to extract the more meaningful ones, what define the difference between these models. From the Information Theory point of view, it has been argued that this reduction request requires to cope with processes related with convergence and divergence of information, because from a computational perspective, the existence of a explicit representation of saliency in a ex-profess map leads to the idea that some class of spatial selection has to be performed during the pre-attentive process of feature detection. In other way, the divergence process form the retinal input to several features maps followed by the convergence in only one saliency map, could not be carried out without obtaining a final representation as complex to interpret as the same original image.

Our proposal to address this dilemma is based in a conjecture presented by Mira and Moreno Diaz [17] about the interrelationship between structures code-meaning at we move towards deeper centers in the central nervous system. This formulation is of vital importance in the structural and functional nervous system theory and is also related with the allocation of meaning to seemingly trivial or complex codes. The basic idea of this approach is that the semantics of a system itself can be distributed between the input-output spaces and the relational structure, so that for the same overall performance, the greater the operational capability of the symbols encoded on variables input, the lower the complexity of the calculations necessary to explain the observed behavior. This suggests a sort of principle of conservation of the amount of complexity contained in the response of a system so that, by increasing the complexity and level of the input data it is reduced the complexity of the rules of decision or operation and vice versa.

In our model the process would act as follows: our starting point is an environmental measured data, which have already been assigned a relative complexity to their extension and intention, that is the number of them and the meaning assigned. In the case of an image captured by the retinal photoreceptors, the extension is great and the intention is minimal. The transformations top-down for the linear filters for obtaining saliency map, help increase intention to assign meanings as those for the detection of relevant features, while the degree of extension is progressively reduced to converge on the saliency map . What is relevant is this interpretation is that completeness remains constant, that is, there is no loss of information throughout the process and the possible loss of data was offset by an increase in the semantic content of the signal.

In this work we consider a computational point of view based on different studies [18] [19], which have demonstrated that the visual process involved from photoreceptors to ganglion cells could be adequately interpreted by means of a set of linear or not linear spatio-temporal filters with center-periphery receptive fields, while the explanations of process performed at a higher levels (NLG and cortical areas) requires hybrid formulations based on cooperative process of algorithmic type. Our proposal for focus attention while preserving information could be applied in both attentional

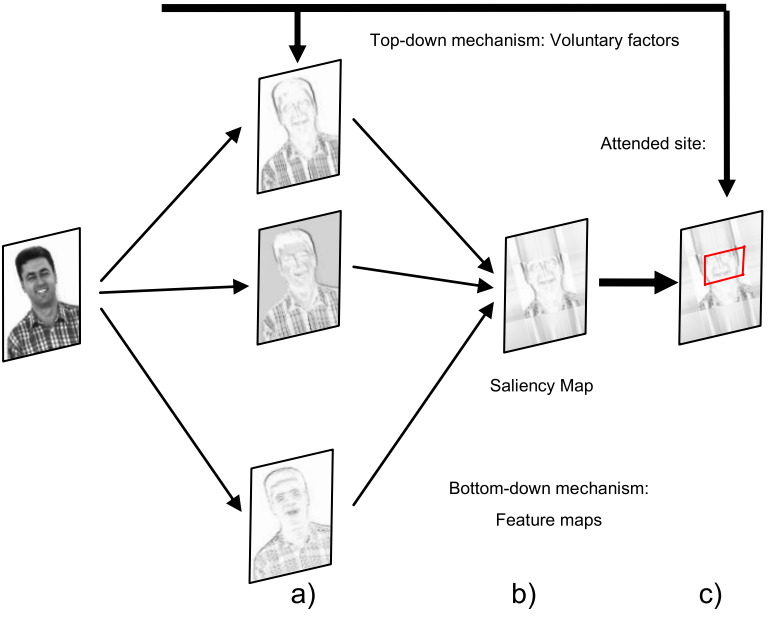


Fig. 1. Proposed model for visual attention: a) Non linear filters acting at the low level visual process and selecting some kind of pop-out characteristic. b) Convergence of information implicit in the saliency map based on cooperative process. c) Selected place carried out by some kind of hybrid formulations of algorithmic type.

strategies context: the first bottom-up strategy based on the fact that at the lowest levels saliency information is systematically extracted from the visual scene, and the second one top-down strategy based on higher levels mechanisms which, in general, involve voluntary factors. This is represented in figure 1.

The saliency map that topographically encodes for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy [5]. With respect to the hybrid formulation based on cooperative process, different works developed in a data filed transform context [14], supported by experimental evidences, has inspired us the application of one particular class of complete transformation, based on the moving average structure. The main idea of this denominated Foveal Transform follows from the fact that in a global complete partition of a data field, it is possible to impose specific receptive fields of our election, from the initial degrees of freedom.

The main characteristic of this transformation is that the higher resolution is concentrated in an area of the image denominated fovea (due to obvious ocular parallelism). In a first approximation this zone of higher resolution was expanded from the geometrical center of the image, while in a revised version of it, the higher resolution domains must be positioned around an image point. This is illustrated specifying over a given retina all the receptive fields form a certain dimension as is shown in figure 2.

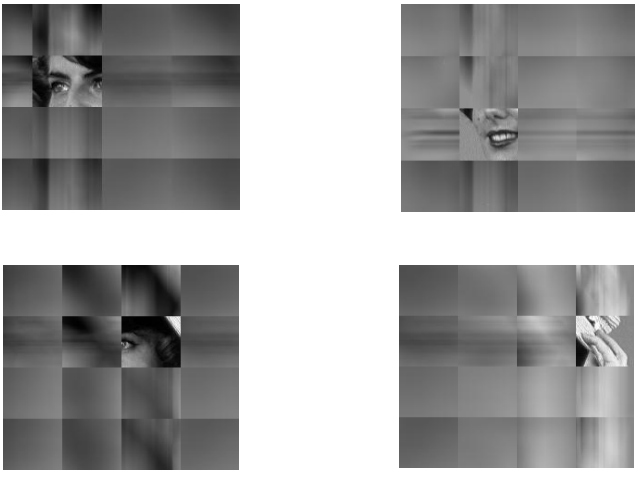


Fig. 2. Illustration of Foveal Transformation in different locations

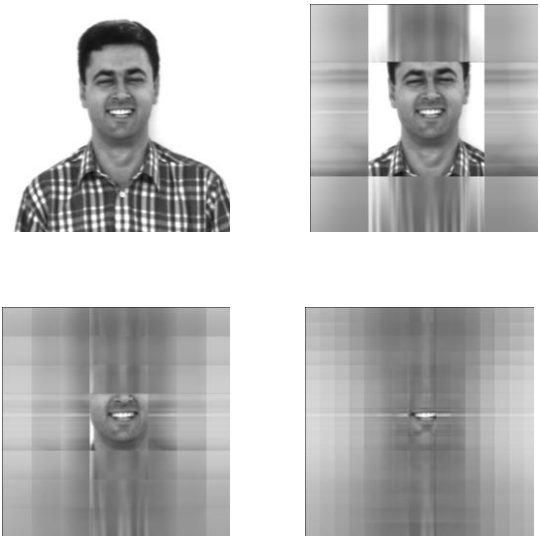


Fig. 3. Illustration of Foveal Transformation with different sizes of fovea

At the same time, it is possible to define the foveal size of a specified position, as it is presented in figure 3. In this way the election of the foveal size would condition, in any way, the resolution level implicit in the multi scale feature extraction, while the position of the fovea would determine the place of interest. The possibility of being able to direct the fovea to any point from the data field is useful to focus attention where the maximum information must be captured and to place the fovea in this image zone, allowing the less relevant information to spread on the parafoveal places.

4 Conclusions

In this work, we have applied the most interesting properties of the layered computing systems to the study of the mechanisms involve in the Visual Attention Process. We have focus our attention in mechanisms such as computing layers, presynaptic inhibition schemes, analysis of the convergence and divergence of information and data fields transformations theory.

From a cybernetics point of view, the above scheme could be considered as a first step in the modelling of the attentional mechanism which governs a great part of the visual process in higher vertebrates. The proposed model allows us to capture or to focus the attention of an observer onto a particular zone of an image and at the same time controlling the size of the fovea, specifying the number of levels required in the transformation , as can be observer in figure 2 and 3 respectively.

In this way, the parafoveal places would extract information from an event, but with certain resolution loss, while once the event of interest is discriminated, the fovea will direct to it to perform a more detailed process in this selected fragment.

Another interesting aspect is the completeness one, being the degree of completeness of a transform a measure of the further recoverability of the original data form the image descriptors selected.

The application of this transform allows us to highlight some feature or object in the image by blurring the rest of it, since we assume that it does not contain any relevant information.

This proposal inspired in Foveal Transformations could be proposed as a first order attention focusing mechanism, being suitable to modelling visual attention due to two main experimental evidences:

1. There are enough evidences that demonstrate that attention process or capacity, besides enhancing the relevant information, also inhibits the potentially distractor one. Different works related with the attentional focus, describe a perifocal zone affected by inhibitory process, where the attentional acuity is worse that in zones remotely situated with respect the center of attention. [12]
2. Our visual system is able to process no selected information [19] and to put on, independently of the attentional control, different process of revision which affects perception and action.

References

1. Leibovic, K.N.: Parallel Processing in Nervous Systems with converging and diverging transmission. In: Biomathematics and Related Computational Problems, pp. 65–72 (1988)
2. Rao, K.R., Ahmed, N.: Orthogonal Transforms for Digital Signal Processing. In: OEEE International Conference on Acoustics, Speech and Signal Processig, pp. 136–140 (1976)
3. Treue, S.: Visual attention: the where, what, how and why of saliency. Current Opinion in Neurobiology 13, 428–432 (2003)
4. James, W.: Princ. Psychol. Harward University Press, Cambridge (1890)
5. Itti, L., Koch, C.: Computacional modelling of visual attention. Nature Reviews. Neuroscience 2(3), 194–203 (2001)

6. Mundhenk, N., Itti, L.: Computacional modelling and exploration of contour integration for visual saliency. Biological Cybernetics 93, 188–212 (2005)
7. Moreno Díaz, R., Santana, O., Rubio, E., Nuñez, A.: Bases de una Teoría de Proceso de datos en la Retina. Biocibernética. Ed. Siglo XXI, pp. 121–131 (1984)
8. Moreno Díaz, R., Rubio, E., Nuñez, A.: A layered model for visual processing in avian retina. Biological Cybernetics 38, 85–89 (1984)
9. Edelman, S.: Representing three-dimensional objects by sets of activities of receptive fields. Biological Cybernetics 70, 37–45 (1993)
10. Weiss, Y., Edelman, S.: Representation of similarity as a goal of early visual processing. Computation in Neural Systems 6, 19–41 (1995)
11. Mira, J., Manjares, A., Ros, S., Delgado, A., Alvarez, J.: Cooperative Organization of Connectivity Patterns and Receptive Fields in the Visual Pathway: Application to Adaptive Thresholding
12. Slotnick, S., Schwarzbach, J., Yantis, S.: Attentional inhibition of visual processing in human striate and extrastriate cortex (2003)
13. Bolivar, O., Candela, S., Muñoz, J.A.: Non linear data Transform in Perceptual Systems. In: Pichler, F., Moreno-Díaz, R. (eds.) EUROCAST 1989. LNCS, vol. 410, pp. 301–309. Springer, Heidelberg (1990)
14. Bolivar Toledo, O., Quevedo, J.C., Muñoz Blanco, J.A.: On the Generation of Representational Spaces in Perceptual Systems. Systems Analysis Modelling Simulation 43(9), 1263–1269 (2001)
15. Rovaris, E.: Images, Sampling and Neuron like Distributed Computing: Towards an Unified Approach. Doctoral Thesis University of Las Palmas de Gran Canaria (1984)
16. Quevedo, J.C., Bolivar, O., Moreno, R.: Cast Methods for Generation of Non-Orthogonal Complete Transforms. In: Albrecht, R., Moreno-Díaz, R., Pichler, F. (eds.) EUROCAST 1995. LNCS, vol. 1030, pp. 447–459. Springer, Heidelberg (1996)
17. Mira, J., Moreno Díaz, R.: Un Marco Teórico para Interpretar la Función Neuronal a Altos Niveles. In: Proc. I Reunion Nacional de Biocibernética. Madrid. Real Academia de Ciencias, pp. 151–178 (1982)
18. López., M.T., Fernández, M.A., Fernandez Caballero, A., Mira, J., Delgado, A.: Dynamic visual attention model in image sequences. Image and Vision Computing 25, 597–613 (2006)
19. Mundhenk, N., Itti, L.: Computacional modelling and exploration of contour integration for visual saliency. Bioliogical Cybernetics 93, 188–212 (2005)

Modelling Fed-Batch Fermentation Processes: An Approach Based on Artificial Neural Networks

Eduardo Valente[1,2], Isabel Rocha[3], and Miguel Rocha[1]

[1] Dep. Informatics / CCTC - University of Minho
Campus de Gualtar, 4710-057 Braga - Portugal
mrocha@di.uminho.pt
[2] Dep. Engenharia Informática, ESTCB
Castelo Branco - Portugal
eduardo@est.ipcb.pt
[3] IBB - Institute for Biotechnology and Bioengineering
Center of Biological Engineering - University of Minho
Campus de Gualtar, 4710-057 Braga - Portugal
irocha@deb.uminho.pt

Summary. Artificial Neural Networks (ANNs) have shown to be powerful tools for solving several problems which, due to their complexity, are extremely difficult to unravel with other methods. Their capabilities of massive parallel processing and learning from the environment make these structures ideal for prediction of nonlinear events. In this work, a set of computational tools are proposed, allowing researchers in Biotechnology to use ANNs for the modelling of fed-batch fermentation processes. The main task is to predict the values of kinetics parameters from the values of a set of state variables. The tools were validated with two case studies, showing the main functionalities of the application.

Keywords: Multilayer Perceptrons, Bioprocess Modelling, Biotechnology, Fed-batch fermentation processes, Bioinformatics.

1 Introduction

The productivity of the manufacturing processes is of paramount concern of any company that wishes to survive in the market. The strong competition leads to the search for new strategies and to the investment in alternatives to traditional methods. Under this scenario, the use of fermentation techniques is steadily growing as a solution to achieve the production of targets with high economic value such as recombinant proteins, antibiotics and amino-acids. Also, they have been replacing traditional processes in areas such as the production of bulk chemicals, given their low energy requirements and environmental costs. However, these processes are typically very complex, involving different transport phenomena, microbial components and biochemical reactions. The nonlinear behavior and time-varying properties limits the application of traditional control and optimization techniques to bioreactors.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 30–39, 2009.
springerlink.com

A fermentation process can be described as any process that produces a specific product through the mass culture of a microorganism [13]. The yield, in this case, means increasing the cell reproduction rates, or the production of compounds from these cells, per unit of time. The factors influencing a fermentation process range from the physical and chemical conditions of the medium (temperature, pH, etc.), sources of energy (glucose, oxygen, light) and nutrients (nitrogen) supplied to the culture. Traditionally, empirical methods based on Monod equations have been adopted for the optimisation of these processes. Each strain of a microorganism has its own specificities and the way it is grown also gives rise to distinct culture behaviour.

The modelling task is essential to achieve the optimization of these processes, since it provides simulated data that are often difficult to measure directly from the culture. One important example is the kinetic behaviour, which depends on factors involving cellular catalysts, intracellular phenomena and characteristics of the population [9], among other complex aspects which makes its collection impossible to be carried out in real-time. Usually, the kinetics is estimated by the measurement of other variables of the process, which in most cases can only be obtained at the end of the experiments. Work in this field has led to the development of tools that attempt to address problems specific to a particular culture, with limited generalization capabilities. Another major problem is the complexity of the models, which use mathematical computations that require lots of processing resources [15, 16].

This work aims to demonstrate that Artificial Neural Networks (ANNs) are adequate to this task, due to their features of massive parallel computing and representation, ability to learn, adaptability and generalization [4]. ANNs, and in particular Multilayer Perceptrons (MLPs) used in this case, are mainly characterized by the topology of connections between neurons, the training algorithm and the activation function [3, 2]. The choice of which ANN will be used to represent a process depends on the characteristics of the problem in question [7]. There are several methods to reach the desired solution, that range from exhaustive search, to local optimization techniques and also the use of Evolutionary Algorithms [5, 14].

The main contribution of this work is the development of computational tools that allow a Biotechnology researcher, with very limited programming skills, to be able to model fermentation processes using ANNs. These will be used to predict the values of kinetic parameters from state variables. A number of tools will be provided to create ANN models from data, to evaluate a set of distinct ANN topologies and also to create artificial data to test ANN models.

2 Modelling Fermentation Processes

A fermentation process can be modelled by a set of Ordinary Differential Equations (ODE), which give the values of state variables at a particular instant. In these equations, there are a number of kinetic parameters to be determined, which is often a quite complex task. Several authors suggested that the best

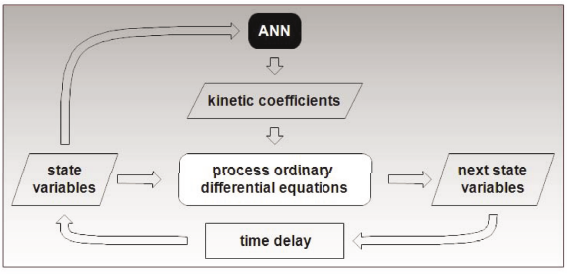


Fig. 1. Grey Box Model structure

approach is to use a hybrid model, where the contribution of various sources of information is merged for the construction of a single model [9, 6, 11, 17]. This approach was adopted as a basis to this work.

The kinetic component is, therefore, modelled using mechanisms that do not require an a priori knowledge of the process. These mechanisms are known as black box models, since only the inputs and the outputs of the system are known, while its internal processing is hidden since it is unknown or too complex. The data used for the construction of the kinetic component of the model can be obtained from real experiments. On the other hand, the structure of the ODEs is typically obtained from literature or biochemical knowledge. This component relies on mechanistic models, which are visible and can be interpreted (white box models). The merging of the two models in a single model leads to a hybrid structure, known as grey box models (Figure 1).

3 Description of the Computational Tools

3.1 Methodology

The software tools were all implemented using Java and an object-oriented approach. The main concerns in development were the modularity to allow code re-usage and the clear separation of the user interfaces and core functionality implementation. The tools are available at http://darwin.di.uminho.pt/bionn.

The computational tools developed in this work contemplate the white box and the grey box modelling strategies. The value of the state variables is always given by ODEs and the numerical simulation is performed using *ODEToJava*, a package of ordinary differential equation solvers [1]. The kinetic behavior can be modelled by two different strategies: heuristic approaches (white box) defined by some expert or taken from literature or the use of ANNs (black box approach). In this last case, a grey box modelling strategy is followed. The tools have been divided into two main groups: the ones related with white box models and the latter related to the grey box models.

3.2 White Box Interface

The White Box interface (Figure 2) includes a number of functionalities in the manipulation and simulation of white box models. In this area, it is possible to produce simulated data from white box heuristic models under different scenarios. The application considers a list of predefined fed-batch processes and models. It is possible to create new processes through the development of a Java source file that defines its behaviour, that is introduced and compiled at run-time. The simulated data generated in this area can be used to train and test ANNs in the grey box interfaces (next section), without the need to collect data from real cultures. The following list of functionalities is available:

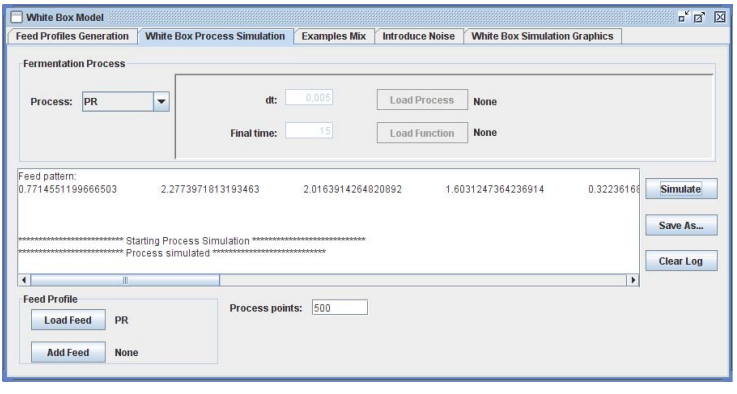


Fig. 2. Functionalities included in the White Box interface

Feeding Profiles Generation. The feeding profiles, i.e. the amount of substrate to provide to the culture at each time, represent the inputs of the simulation. Through this panel multiple profiles can be created with constant, linear, random, saw wave or sinus functions. Using different feeds to create data allows a broader spectrum of training examples.

White Box Process Simulation. This simulation uses heuristic models for generating the kinetics coefficients, which are then introduced into the ODE models. For each feeding profile, a distinct simulation can be conducted resulting in the values for state variables and kinetic parameters over time.

Mixing Examples. A number of files with simulations can be handled, and used to create new files through composition, cutting, interpolation or sampling data. These can create more comprehensive sets of examples or decrease the size of some collections without losing its generality.

Introducing Noise. The introduction of noise on data serves to create data sets that increase the generalization abilities of the ANN model. This can be useful when the set of available data is not rich enough to provide a good training process for the ANN.

White Box Simulation Graphs. The graphs allow for a better visualization of the data generated by the mathematical model. The preview is done in two separate graphs for the state variables and kinetics coefficients. It also allows visual comparisons between results of different simulations.

3.3 Grey Box Interface

The tools related to the grey box models include the training and evaluation of ANNs from experimental or simulated data and also the calculation of the model kinetic variables to integrate in the overall simulation (Figure 3). The following list of functionalities is available in this area:

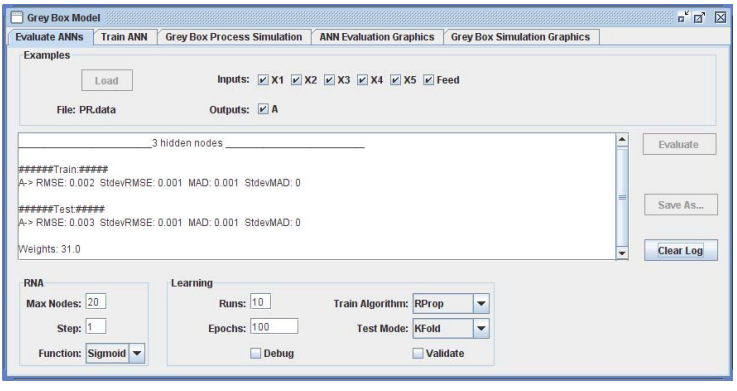


Fig. 3. Functionalities included in the Grey Box interface

Evaluate ANNs. This tool can assess various configurations of ANNs, varying the input and output variables, the number of nodes in the hidden layer (all ANNs are MLPs with one hidden layer, completely connected), training algorithm and the number of training epochs. Each model is evaluated by its training error and generalization error in a validation set. It is thus possible to determine which parameter combinations will lead to more accurate models.

ANN training. An ANN with a given configuration (typically the one that resulted from the above study) can be trained. It then can be used to predict the kinetic variables, integrated within the grey box model.

Grey Box Process Simulation. Enables the simulation of the fermentation process. Trained ANNs calculate the kinetic coefficients from state variables and ODEs simulate the mass-balance equations that determine the evolution of the values of the state variables.

ANN Evaluation Graphs. Show the errors from the evaluation of ANN configurations. They provide information on the settings, data and variables that were used in the training. The graphs allow a better interpretation of the error curves (both in training and validation sets).

Grey Box Simulation Graphs. Show data resulting from grey box model simulations. They show the evolution for each state variable and kinetic coefficient. It is also possible to compare two simulations and thus examine, for example, the difference between the curves obtained in the simulation model with grey box and the ones obtained with white box models.

4 Case Studies

4.1 PR Process

This process represents a *Saccharomyces cerevisiae* culture and was studied by Park and Ramirez (PR) [10]. The model equations (ODEs) are the following:

$$\frac{dx_1}{dt} = \frac{4.75A(x_2 - x_1)}{0.12 + A} - \frac{ux_1}{x_5} \tag{1}$$

$$\frac{dx_2}{dt} = \frac{x_3 x_4 e^{-5x_4}}{0.1 + x_4} - \frac{ux_2}{x_5} \tag{2}$$

$$\frac{dx_3}{dt} = (A - \frac{u}{x_5})x_3 \tag{3}$$

$$\frac{dx_4}{dt} = -7.3Ax_3 - \frac{u(x_4 - 20)}{x_5} \tag{4}$$

$$\frac{dx_5}{dt} = u \tag{5}$$

where x_1, x_2, x_3 and x_4 are the concentrations of secreted protein (units/L), total protein (units/l), cells (g/l) and substrate (g/l) respectively; x_5 is the fermenter's volume (l) and u the feed rate (l/h). The specific growth A (h^{-1}) is the only kinetic parameter and follows substrate inhibition kinetics:

$$A = \frac{21.87x_4}{(x_4 + 0.4)(x_4 + 62.5)} \tag{6}$$

4.2 Ecoli Process

This case study consists of an *Escherichia coli* culture. The kinetic behaviour is complex, presenting various coefficients and kinetic points of discontinuity which make this model a real challenge for the ANNs. The model is given by the following equations [12]:

$$\frac{dX}{dt} = (\mu_1 + \mu_2 + \mu_3)X - DX \tag{7}$$

$$\frac{dS}{dt} = q_s X + \frac{F_{in,S} S_{in}}{W} - DS \tag{8}$$

$$\frac{dA}{dt} = (k_3\mu_2 - k_4\mu_3)X - DA \tag{9}$$

$$\frac{dO}{dt} = (-k_5\mu_1 - k_6\mu_2 - k_7\mu_3)X + OTR - DO \tag{10}$$

$$\frac{dC}{dt} = (k_8\mu_1 + k_9\mu_2 + k_{10}\mu_3)X - CTR - DC \tag{11}$$

$$\frac{dW}{dt} \simeq F_{in,S} \tag{12}$$

being D the dilution rate, $F_{in,S}$ the substrate feeding rate (in kg/h), W the fermentation weight (in kg), OTR the oxygen transfer rate and CTR the carbon dioxide transfer rate.

The kinetic behavior is expressed in four variables: the rates μ_1 to μ_3 and q_s. A set of heuristic rules to calculate these values from state variables was defined in [12] and it is not shown here given its complexity. However, since it involves conditional branches it is discontinuous and also nonlinear.

4.3 Methodology

Using the previous white box models, a set of simulations was conducted with distinct feeding profiles. These were then be used to perform ANN evaluations and to test grey box models. To achieve this objective, it was necessary to follow a structured methodology to withdraw the best use of the case study. Overall, this approach has been structured into five steps: model selection; generation and selection of cases for training, selection of architecture and parameters; ANN training and simulation of the fermentation process.

5 Results

5.1 PR

For this case study, 22 distinct feed profiles were used: 4 random, 4 linear and 3 constants profiles and also included 11 optimum feed profiles (obtained by optimization with Evolutionary Algorithms [8]). All feedings have 31 values, which were linearly interpolated. Each of the profiles was used to generate a data file and all files were merged. Afterwards, the final training data set was obtained by sampling only 150 examples. This set of examples was used to test different ANN configurations by using 10-fold cross validation.

The first set of tests was conducted using only one state variable (1 ANN input). Then, other tests were performed in which all except one state variable were used. The results were compared, taking as a basis for comparison the case in which all state variables available were used as inputs. The comparative results showed that the best configuration would be the use of only the variable X4, a result that was expected given the expression used to compute A.

Regarding the number of hidden nodes, it was shown that the use of one intermediate node would already be viable to predict the kinetics accurately. However, 6 hidden nodes provided better generalization errors. So, an ANN with this configuration has been trained and used to simulate the process. In Figure 4 the results of the simulation to the process with this ANN are shown.

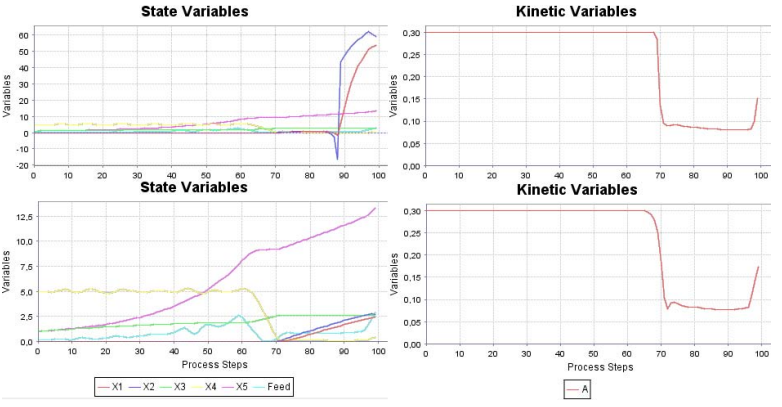


Fig. 4. Top: Simulation of the process PR (White Box). Bottom: Simulation of the process model with grey box model.

5.2 Ecoli

For this case study, 21 different feed profiles were used: 5 random, 5 linear, 4 constants and 7 optimum profiles. All feedings contained 26 values and each of the profiles was used to generate the data files. The method of selection of the state variables that influence the kinetic coefficients of the process was similar to the one used for PR. In this case, the conclusion was that the best configuration would be the use of variables OTR, CTR, X, A and F. This configuration was used to simulate the process using ANNs to compute the kinetic parameters, obtaining the result shown in Figure 5:

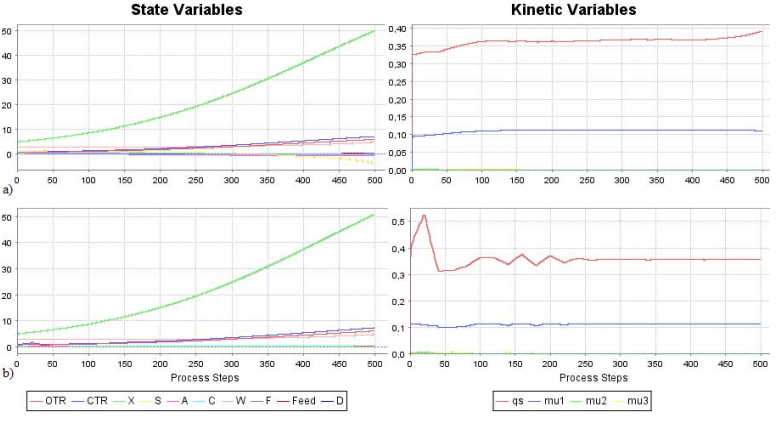


Fig. 5. Top: Simulation of the process Ecoli (White Box). Bottom: Simulation of the process with grey box model.

6 Conclusions and Further Work

A set of computational tools were developed to aid in the modelling of fed-batch fermentation processes. These can be used to handle data from real experiments or using simulated data. It is thus possible to shape a process without the need to know the mathematical description of its mechanism, since learning is made directly from examples created from experimental data, showing the potential of the application of ANNs in this type of problems.

It is possible to introduce new fermentation processes at run-time, enabling the application for a multitude of cases. It not only seeks to provide a solution to a specific case, but serves as a platform for a layman in ANNs to evaluate and use multiple models for each culture. This is also possible since user interfaces are implemented in a simple and intuitive way. The modularity of the application makes it scalable; the main modules can easily be used for future work. The core implementation of the functionalities is detached from the user interface, to make code understanding and re-use easier.

The implementation of several graphs allows making visual comparisons between results, providing easier analysis of the factors influencing the course of the process. Thus, it is possible to identify potential optimizations that can be made, such as to change the feed profile. These optimizations can be tested within the tool, thereby avoiding some real experiments that brings, necessarily, increased costs for research.

In the future, a major aim is to improve the capabilities of the application with new tools. A major concern is to create an interface to make the introduction of new processes easier, without having to write Java code.

Acknowledgments

This work was supported by the Portuguese FCT under project POSC/EIA/59899/2004.

References

1. Ascher, S., Ruuth.: Implicit-explicit runge-kutta methods for time-dependent partial differential equations. Applied Numerical Mathematics 25, 151–167 (1997)
2. Coulman, G.A., Stieber, R.W., Gerhardt, P.: Dialysis Continuous Process for Ammonium-Lactate Fermentation of Whey: Mathematical Model and Computer Simulation. American Society for Microbiology (1977)
3. Haykin, S.: Neural Networks - A Compreensive Foundation, 2nd edn. Prentice-Hall, New Jersey (1999)
4. Jain, A.K., Mao, J., Mohiuddin, K.M.: Artificial Neural Networks: A Tutorial. IEE (1996)
5. Lednick, P., Mészàros, A.: Neural Network Modeling in Optimization of Continuous Fermentation Process. Bioprocess Engineering 18, 427–432 (1998)
6. Lee, D.S., Park, J.M.: Neural Network Modeling for On-line Estimation of Nutrient Dynamics in a Sequentially-operated Batch Reactor. Journal of Biotechnology 75, 229–239 (1999)

7. Levisauskas, D., Tekorius, T.: Model-Based Optimization of Fed-Batch Fermentation Processes Using Predetermined Type Feed-Rate Time Profiles. A Comparative Study. In: ITC (2005)
8. Mendes, R., Rocha, M., Rocha, I., Ferreira, E.C.: A Comparison of Algorithms for the Optimization of Fermentation Processes. In: Proceedings of the 2006 IEEE Conference on Evolutionary Computation, pp. 7371–7378. IEEE Computer Society Press, Los Alamitos (2006)
9. Oliveira, R.: Combining First Principles Modelling and Artificial Neural Networks: A General Framework. Computers and Chemical Engineering 28, 755–766 (2004)
10. Park, S., Ramirez, W.F.: Optimal Production of Secreted Protein in Fed-batch Reactors. AIChE J. 34(9), 1550–1558 (1988)
11. Peres, J., Oliveira, R., Azevedo, S.F.: Knowledge Based Modular Networks for Process Modelling and Control. Computers and Chemical Engineering 25, 783–791 (2001)
12. Rocha, I.: Model-based strategies for computer-aided operation of recombinant E. coli fermentation. PhD thesis, Universidade do Minho (2003)
13. Stanbury, P.F., Whitaker, A.: Principles of Fermentation Technology. Pergamon Press, Oxford (1984)
14. Taylor, B.J.: Methods and Procedures for the Verification and Validation of Artificial Neural Networks. Springer, Heidelberg (2006)
15. Veloso, A.C., Rocha, I., Ferreira, E.C.: On-Line Estimation of Biomass in an E. Coli Fed-Batch Fermentation. In: Enpromer (2005)
16. Zheng, Y., Gu, T.: Analytical Solutions to a Model for the Startup Period of Fixed-Bed Reactors. Elsevier Science (1996)
17. Zuo, K., Wu, W.T.: Semi-realtime Optimization and Control of a Fed-batch Fermentation System. Computers and Chemical Engineering 24, 1105–1109 (2000)

New Principles and Adequate Control Methods for Insulin Dosage in Case of Diabetes

Levente Kovács

Departament of Control Engineering and Information Technology, Budapest University of Technology and Economics, Magyar tudosok krt. 2, Budapest, Hungary
lkovacs@iit.bme.hu

Abstract. The paper is a short summary of the author's PhD dissertation with the same title [1], submitted at the Budapest University of Technology and Economics in November 2007. The thesis is a multidisciplinary work including physiological modelling and control, control engineering and informatics and the new scientific results are structured on three parts: modelling concepts of Type I diabetes, robust control methods for optimal insulin dosage and symbolic computation-based robust algorithms using *Mathematica*.

Keywords: diabetes mellitus, glucose-insulin control, minimal-model, Sorensen-model, robust control, LPV control, *Mathematica*.

1 Introduction

Diabetes mellitus is one of the most serious diseases which need to be artificially regulated. The statistics of the World Health Organization (WHO) predate an increase of adult diabetes population from 4% (in 2000, meaning 171 million people) to 5,4% (366 million worldwide) by the year 2030 [2]. This warns that diabetes could be the "disease of the future", especially in the developing countries (due to the stress and the unhealthy lifestyle).

In many biomedical systems, external controller provides the necessary input, because the human body could not ensure it. The outer control might be partially or fully automatized. The self-regulation has several strict requirements, but once has been designed it permits not only to facilitate the patient's life suffering from the disease, but also to optimize (if necessary) the amount of the used dosage.

The blood-glucose control is one of the most difficult control problems to be solved in biomedical engineering. One of the main reasons is that patients are extremely diverse in their dynamics and in addition their characteristics are time varying. Due to the inexistence of an outer control loop, replacing the partially or totally deficient blood-glucose-control system of the human body, patients are regulating their glucose level manually. Based on the measured glucose levels (obtained from extracted blood samples), they decide on their own what is the necessary insulin dosage to be injected. Although this process is supervised, mishandled situations often appear. Hyper- (deviation over the basal glucose level) and hypoglycemia (deviation under the basal glucose level) are both dangerous cases, but on short term the latter is more dangerous, leading for example to coma.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 40–44, 2009.
springerlink.com

Starting from the Seventies lot of researchers investigate the problem of the glucose-insulin interaction and control. The closed-loop glucose regulation as it was several times formulated [3, 4, 5], requires three components: glucose sensor, insulin pump, and a control algorithm, which based on the glucose measurements, is able to determine the necessary insulin dosage.

The author's PhD dissertation [1], focused on the last component and analyzed robust control aspects of optimal insulin dosage for Type I diabetes patients.

To design an appropriate control, an adequate model is necessary. In the last decades several models appeared for Type I diabetes patients. The mostly used and also the simplest one proved to be the minimal model of Bergman [6] and its extension, the three-state minimal model [7]. However, the simplicity of the model proved to be its disadvantage too, while in its formulation a lot of components of the glucose-insulin interaction were neglected. Therefore, the model is valid only for Type I diabetes patients under intensive care. The dynamic characteristics of the model are created by artificially dosed glucose input. As a result, the model can simulate only a 3 hours period. Furthermore, it was demonstrated, that the model control possibilities are limited, while it is very sensitive to its parameters variance.

Henceforward, extensions of this minimal model have been proposed [8, 9, 10, 11], trying to capture the changes in patients' dynamics, particularly with respect to insulin sensitivity, Also with respect to the meal composition, minimal model extensions were created [12, 13]. In the PhD thesis cited by the current paper [1], the author used the modified minimal model of Bergman proposed by [11], as well as the extended three-state minimal model [7].

Beside the Bergman-model other more general, but more complicated models appeared in the literature [14, 15]. The most complex one proved to be the 19th order Sorensen-model [15]. Even if the model describes in the most exact way the human blood glucose dynamics, its complexity made it to be rarely used in research problems. Nowadays, it is again more often investigated (due to its general validity), therefore the second model considered by the author in his PhD thesis is the Sorensen-model.

Regarding the applied control strategies, the palette is very wide [16]. Starting from classical control strategies (ex. PID control [17]) to soft-computing techniques (ex. neuro-fuzzy methods [18]), adaptive [11], model predictive [3, 19], or even robust H_∞ control were already applied [4, 5]. However, due to the excessive sensitivity of the model parameters (the control methods were applied mostly on the Bergman minimal model), the designed controllers were true only for one (or in best way for few) patient(s).

As a result, investigations demonstrated [3, 5], that even if the best way to approach the problem is to consider the system model and the applied control technique together, if high level of performance is desired, a low complexity control (like PID) is not effective. Therefore, the literature has oriented in two directions: adaptive control and robust control techniques.

The advantage of the adaptive control is the retuning possibility of the controller even in its working conditions. However, its disadvantage appeared if the complexity of the diabetes model was grown. Robust control adjusted the disadvantages of the adaptive control technique, but the designing steps are more difficult. Due to the fact that the literature has clearly presented the adaptive control possibilities of the glucose-insulin control, the PhD dissertation of the author [1], focuses on the robust control methodology.

2 New Modelling Concepts for Type I Diabetes

The proposed modeling formalisms cover the analytical investigation of the high complexity Sorensen-model and the extension of the modified Bergman minimal model. In this way, the proposed approximations are indicating numerical algorithmization for complex optimal control strategies while cover a bigger diabetes population.

For the extension of the modified minimal model, an internal insulin device was proposed. In this way, without damaging the simple structure of the Bergman model it was possible to model not only the Type I intensive care situation, but also the physiological variation of the interstitial insulin.

In case of the Sorensen-model, for an easier handling, inside the physiological boundaries, an LPV (Linear Parameter Varying) modeling formalism was proposed. In this way the model is possible to be reduced to a corresponding degree and consequently to ease the control possibilities and the applicability of the Sorensen-model.

3 Robust Control Methods for Optimal Insulin Dosage in Case of Type I Diabetic Patients

The proposed robust control methods for insulin dosage were structured on the two considered models.

Firstly, the modified minimal model of Bergman was investigated. The mini-max control method was developed comparing it with the classical LQ method. Furthermore, using the μ-synthesis method, parameter uncertainty was taken into account, which supplements the H_∞ method in guaranteeing the robust performance requirements. Moreover, with suitable parameterization, a quasi-Affine Linear Parameter Varying system-set have been defined and exploiting this result a (nonlinear) controller was designed ensuring quadratic stability.

Regarding the Sorensen-model, using the normoglycaemic insulin input, the high complexity Sorensen-model was parameterized and described with politopic LTI (Linear Time Invariant) systems. With the so built LPV model a corresponding controller using induced L_2 norm minimization was designed. Finally, with the nonlinear (LPV) controller the nonlinear Sorensen-model was controlled guaranteeing γ performance level.

4 Symbolic Computation-Based Robust Algorithms with *Mathematica*

To ease the applicability of the applied robust methods, user-friendly symbolic algorithms were developed under *Mathematica*, which help the introduction of the so developed insulin dosage algorithms in therapeutics as well as in education.

Firstly, the extended LQ (minimax) method was symbolically implemented under *Mathematica*. It was shown how MATLAB selects its own solution (from the two resulting solutions), and a general formula was determined for the worst-case result in case of the modified minimal model of Bergman.

Secondly, regarding the modified minimal model of Bergman it was shown, that the applicability of the minimax method has practical limitations. Therefore, for the modified minimal model of Bergman a solution was proposed, using Gröbner-bases, which spans these limitations. In this way, even if the worst case solution cannot be achieved, it is possible to obtain a better solution than the classical LQ one.

Finally, the graphical interpretation of the H_∞ method implemented under *Mathematica* uses a requirement envelope. For the disturbance rejection criteria the author formulated and extended the requirement envelope's criterion-set with another criterion. The correctness of this „plus" criterion was demonstrated on the extended minimal model of Bergman, and it was compared with literature results. It was presented that the constant used in the proposed plus criteria can be connected with the sensor noise weighting function used under MATLAB.

Acknowledgments. This work was supported in part by Hungarian National Scientific Research Foundation, Grants No. OTKA T69055 and RET-04/2004.

References

1. Kovács, L.: New principles and adequte control methods for insulin optimization in case of Type I diabetes mellitus, PhD Thesis (in Hungarian), Budapest University of Technology and Economics, Hungary (2007)
2. Wild, S., Roglic, G., Green, A., Sicree, R., King, H.: Global Prevalence of Diabetes - Estimates for the year 2000 and projections for 2030. Diabetes Care 27(5), 1047–1053 (2004)
3. Hernjak, N., Doyle III, F.J.: Glucose control Design Using Nonlinearity Assessment Techniques. AIChE Journal 51(2), 544–554 (2005)
4. Parker, R.S., Doyle III, F.J., Ward, J.H., Peppas, N.A.: Robust H∞ Glucose Control in Diabetes Using a Physiological Model. AIChE Journal 46(12), 2537–2549 (2000)
5. Ruiz-Velazquez, E., Femat, R., Campos-Delgado, D.U.: Blood glucose control for type I diabetes mellitus: A robust tracking H∞ problem. Elsevier Control Engineering Practice 12, 1179–1195 (2004)
6. Bergman, B.N., Ider, Y.Z., Bowden, C.R., Cobelli, C.: Quantitive estimation of insulin sensitivity. American Journal of Physiology 236, 667–677 (1979)
7. Bergman, R.N., Philips, L.S., Cobelli, C.: Physiologic evaluation of factors controlling glucose tolerance in man. Journal of Clinical Investigation 68, 1456–1467 (1981)
8. Fernandez, M., Acosta, D., Villasana, M., Streja, D.: Enhancing Parameter Precision and the Minimal Modeling Approach in Type I Diabetes. In: Proceedings of 26th IEEE EMBS Annual International Conference, San Francisco, USA, pp. 797–800 (2004)
9. Morris, H.C., O'Reilly, B., Streja, D.: A New Biphasic Minimal Model. In: Proceedings of 26th IEEE EMBS Annual International Conference, San Francisco, USA, pp. 782–785 (2004)
10. de Gaetano, A., Arino, O.: Some considerations on the mathematical modeling of the Intra-Venous Glucose Tolerance Test. Journal of Mathematical Biology 40, 136–168 (2000)
11. Juhász Cs, Medical Application of Adaptive Control, Supporting Insulin-Therapy in case of Diabetes Mellitus. PhD dissertation (in Hungarian), Budapest University of Technology and Economics, Hungary (1997)
12. Anirban, R., Parker, R.S.: Mixed Meal Modeling and Disturbance Rejection in Type I Diabetic Patients. In: Proceedings of the 28th IEEE EMBS Annual International Conference, New York City, USA, pp. 323–326 (2006)

13. Roy, A., Parker, R.S.: Dynamic Modeling of Free Fatty Acid, Glucose, and Insulin: An Extended "Minimal Model". Diabetes Technology & Therapeutics 8, 617–626 (2006)
14. Hovorka, R., Shojaee-Moradie, F., Carroll, P.V., Chassin, L.J., Gowrie, I.J., Jackson, N.C., Tudor, R.S., Umpleby, A.M., Jones, R.H.: Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT. American Journal Physiology Endocrinology Metabolism 282, 992–1007 (2002)
15. Sorensen, J.T.: A Physiologic Model of Glucose Metabolism is Man and Its use to Design and Assess Improved Insulin Therapies for Diabetes. PhD dissertation, Massachusetts Institute of Technology, USA (1985)
16. Parker, R.S., Doyle III, F.J., Peppas, N.A.: The Intravenous Route to Blood Glucose Control. A Review of Control Algorithms for Noninvasive Monitoring and Regulation in Type I Diabetic Patients. IEEE Engineering in Medicine and Biology, 65–73 (2001)
17. Chee, F., Fernando, T.L., Savkin, A.V., van Heeden, V.: Expert PID Control System for Blood Glucose Control in Critically Ill Patients. IEEE Transactions on Information Technology in Biomedicine 7(4), 419–425 (2003)
18. Dazzi, D., Taddei, F., Gavarini, A., Uggeri, E., Negro, R., Pezzarossa, A.: The control of blood glucose in the critical diabetic patient: A neuro-fuzzy method. Journal of Diabetes and Its Complications 15, 80–87 (2001)
19. Hovorka, R., Canonico, V., Chassin, L.J., Haueter, U., Massi-Benedetti, M., Orsini Federici, M., Pieber, T.R., Schaller, H.C., Schaupp, L., Vering, T., Wilinska, M.E.: Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. Physiological measurement 25, 905–920 (2004)

A Framework for CBR Development and Experimentation with Application to Medical Diagnosis

Beatriz López, Pablo Gay, Albert Pla, and Carles Pous

Universitat de Girona, Campus Montilivi, edifice P4, 17071 Girona
beatriz.lopez@udg.edu, {pgay,apla}@eia.udg.edu, carles.pous@udg.edu

Summary. Most of the data mining techniques focus on classification tasks and so several tools have been designed to help engineers and physician in building new classification systems. However, most of the tools does not take into account the application domain. In this paper we present a case-based reasoning (CBR) tool with the aim to support the development and experimentation of a particular medical classification task: diagnosis. Our tool, eXiT*CBR provides navigation facilities to the engineers and physician across several experimentation results, assures experiment reproducibility and incorporates functionalities for independizing data, if required. The use of eXiT*CBR is illustrated with a Breast Cancer diagnosis system we are currently developing.

1 Introduction

There is an increasing interest on the use of data mining techniques in the Life Sciences [4, 17, 11, 12]. Particularly, learning techniques for classification tasks are widely present in several fields. As an example, the gene expression settings can be cited, when we might want to predict the unknown functions of certain genes, given a group of genes whose functional classes are already known [16]. We have another example in medical diagnosis, where there may exists many cases that correspond to several diseases, together with their associated symptoms. Classification tools can offer a second opinion to the physicians with a potentially low cost, noninvasive solution to improving the diagnosis [6]. Thus, given a set of sample data, the objective of learning classification techniques is to build such a method that enables the classification of new data successfully, with the corresponding error rates as low as possible.

Our research concerns Case-Based Reasoning (CBR). Case-based reasoning is an approach to problem solving and learning based on examples (past experiences). It consist of four stages, that are repeated for each new situation: Retrieval, Reuse, Revise, and Retain. *Retrieval* consists on seeking for past situation or situations similar to the new one . *Reuse* is the second step and it uses the extracted cases to propose a possible solution. *Revise* the solution proposed in the reuse phase is often a human expert. Finally, in the *Retain* it has to be

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 45–54, 2009.
springerlink.com

decided if it is useful to keep the new situation in the case base in order to help on the future diagnosis of new situation.

Each particular domain requires the selection of the appropriate techniques for each phase and the appropriate tuning of the different parameters. Several tools have been designed in order to guide the development as for example cf [3], CBR* [10] and CBR Shell [2]. They are general purpose tools regarding CBR, so they can be used for developing a cancer diagnosis system as well as an electricity fault diagnosis system. When designing a new tool, however, it is important to identify the target of the user, since each discipline has its own particularities that establish some requirements on the tool. Particularly, the field of Medical diagnoses requires to provide a user-friendly interface, result visualization based on ROC curves, and experiment reproducibility. First, a user-friendly interface allows physicians and engineers to work side by side when defining a new system. This requirement is common to most of the previous techniques. Second, ROC curves are the kind of visualization tools with which physicians usually work and helps them in interpreting the obtained results. ROC (Receiver Operator Characteristics) curves depict the tradeoff between hit rate and false alarm rate [8]. Finally, experiment repetition is often a critical but necessary process to assure the definite parameters that provide the best results. And what is more important, assures reproducibility, so that the results obtained by another team working independently agree given the same method and identical test material. From our knowledge, nowadays there is no tool supporting these two features. This two features provides, in addition to a framework for developing new systems, a platform that helps in the CBR experimentation.

In this paper we present a new tool, eXiT*CBR, with the aim of providing support to the development and experimentation of case-based classification in medicine, that satisfy the above requirements. This paper is an extension of the work initiated in [19]. In this paper, we provide and explain the different modules of the architecture of our tool, as well as the common data representation required for any application. We have also extended the system with additional functionalities regarding case independency. Case independency is assumed by most of the data mining methods, but should be specifically tested when dealing with medical applications, since most of the times this assumption does not hold.

This paper is organized as follows. First, we introduce the architecture of our system. Next, we briefly describe the functionalities of eXiT*CBR. We continue by giving some details of the breast cancer diagnostic application being developed with the tool. Afterwards, we relate our work with previous ones of the state of the art. We end the paper with some conclusions.

2 eXiT*CBR Framework

The aim of our tool is to help engineers and physicians to achieve the appropriate CBR system for medical diagnosis, given the data they have and based on an empirical approach. For this purpose, our architecture follows a modular approach based on the different phases required by any CBR system (see Figure 1).

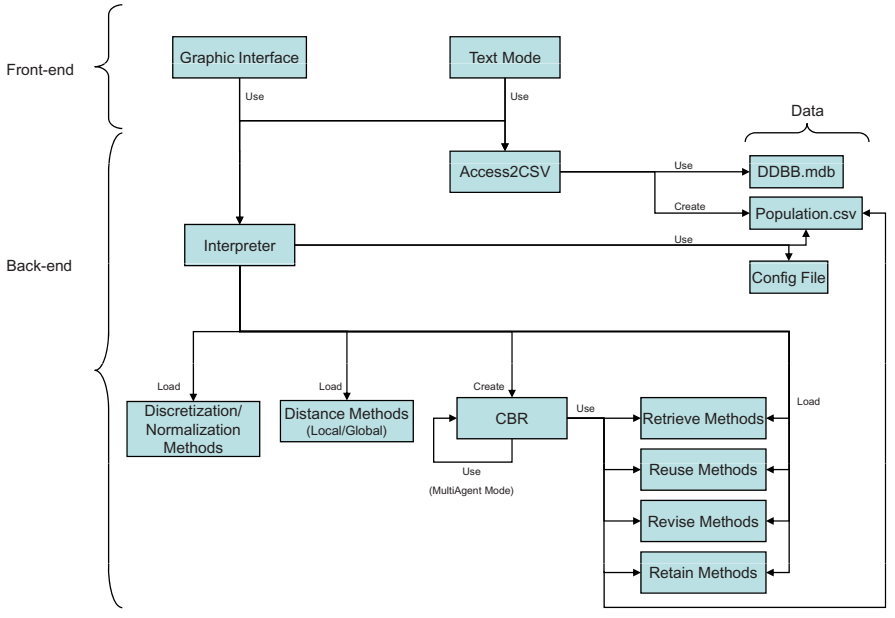


Fig. 1. Tool architecture

Particularly, we distinguish the following modules: Pre-process, Retrieve, Reuse, Revise, Retain, Experimentation, Post-process. Thanks to the object oriented implementation of the architecture, that provides modularity, reusability and extensibility properties, any module is implemented as a generic class. So when a new method is required and not provided in the system, it can be added as a particular instance of the generic class.

In order to handle the data by any method, a common representation is required. In eXiT*CBR we have chosen a plain csv file. The structure of the file is the following:

- First row corresponds to the attribute descriptions (for example, "Age when the first kid was born")
- Second row corresponds to the attribute name (usually in a cryptic form, as for example, "age1stborn").
- Third row corresponds to the attribute type (-1 ignore, 0 discrete, 1 numerical, 2 textual, 3 data)
- Fourth row corresponds to the attribute weight.

We believe that this simply plain representation covers most of the data used in Medical applications[1]. The advantage is that this representation is easy to

[1] In fact, physicians are usually gathering data in Microsoft excel or Spss files, with a similar format.

manage and general enough to be used by any of the current techniques of CBR (mainly, distance functions).

2.1 Pre-processing Module

As preprocessing steps we can have mainly three kind of process: discretization, normalization and feature selection. Discretization methods deal with numerical to categorical data conversion [15]. Normalization methods assures that numerical values are all comparable in the [0,1] interval. Finally, feature selection methods determine which of the available features are relevant for the application. Usually relevant and irrelevant features can be labeled with weights in a weighted learning process, so that relevant features have high weights, while irrelevant features have low weights assigned [13, 15].

Note, however, that in order to learn which features are relevant, other Data Mining or Computing techniques should be considered as a previous step of the CBR, not considered as pre-processing. For this purpose, eXiT*CBR admits plug in of any Data Mining technique in the same interface.

2.2 Retrieve Module

The key issue in the retrieve phase is the definition of the similarity measure. There are local and global similarity measures. On one hand, local similarity measure concern the comparison of two data. There can be as many kind as local similarity measures as type of operands available. For example, the Euclidean distance is the most often proposed measure to handle numeric data, while the Hamming distance is set for categorical (discrete) data [22]. Other local similarity measures regarding data trees (to handle, for example, inheritance information), series, and dates can also be used. On the other hand, global similarity measures is related to the combination of the different local similarity measures. An example of them is the weighted average, but much other can be considered [20].

Other important methods in this phase are the ones related to unknown values: missing completely at random (MCAR), missing at random (MCR), and not missing at random (NMAR) [23]. MCAR is when the probability of missing a value is the same for all attributes, MCR is when the probability of missing a value is only dependent on other attribute, and NMAR is when the probability of missing a value is also dependent on the value of the missing attribute. Missing values are cost sensitive, as explained in [23], and widely present in Medical applications.

Finally, the selection method employed should also be determined: (e.g. k-nearest neighbor).

2.3 Reuse Module

One of the principal limitation of CBR in medical diagnosis is the reuse phase [14]. The majority of medical-CBR rely on suggesting past solutions without further adaptation process. Recent works as [5] propose a probabilistic approach;

however, the Bilska-Wolak method has been applied with a few number of features, and much more research effort should be done in order to deploy it in real environments.

2.4 Revise Module

As a revise phase, most of the current medical-CBR relies on human feedback. Up to now, no further alternatives are open. However, in other environments, simulators are also possible [9].

2.5 Retain Module

Once the solution proposed for the new presented case is revised, a decision has to be made concerning to the necessity of retaining the new case. This decision will depend on whether the cases that we already have, produce a correct solution or not. Since CBR system's core is a base of cases that can be very large, it has a lot in common with the data mining methodologies for data processing. When the case base grows, it is necessary to have a good maintenance policy. This means that we have to delete, add or modify cases in order to keep the system performing well. In this stage Instance based Learning algorithms such as IB3 ([1]) or DROP4 ([21]) can be applied to help on the decision of just keeping the new case, forgetting it, or keeping it in expenses of another cases deletion.

2.6 Experimentation Module

There are three main experimentation parameters: experimentation measures, experimentation methodology, and visualization result method. First, experimentation measures concerns the kind of measure that physicians are interested on. Others as recall, precision, success, failure and much more can be considered, although specificity and sensibility are most commonly used in the medical domain [8]. Second, the experimentation method can be any statistically supported methodology, being the stratified cross-validation technique the one that has proved to be the most adequate when few cases are available, as in medical applications happens [7]. Finally, the usual way of visualizing results in medical applications is by means of ROC curves [5]; even so, other alternatives such as cost plots, and others can also be used.

2.7 Post-processing Module

Current trends on data mining systems should consider different use of the models learned, and consistently, different kinds of post-processing steps should be taken into account. Mainly, we distinguish between off-line and on-line modes. In an off-line mode, the system is being developed, and thus several batch processes and validation procedures should be applied according to the experimentation parameters. The answer of the system in this context are the visualization plots. On the other hand, in the on-line mode the system cooperates with other systems in the solution of a given problem. There could be several approaches based on ensemble learning techniques, including distributed and agent-based approaches [18].

3 eXiT*CBR Functionalities

The aim of our tool is to help engineers and physicians to achieve the appropriate CBR system for Medical diagnosis, given the data they have, based on an empirical approach. For this purpose, our architecture offers the following set of functionalities regarding the user interaction:

1. **Edit configuration:** the user can define a configuration file in a friendly user manner, in which all the available techniques are prompted out in pop-up menus.
2. **Data independization:** Machine learning methods in general, and CBR in particular, assume that examples are independent. However, medical data can not always satisfy this constraint. For example, in a breast cancer data, two members of the same family can be also in the same dataset. This clearly dependency situation can be removed, if family relationship are identifiable in the dataset. Of course, other kind of hidden dependencies are more difficult to handle.
3. **Data conversion:** transforms any original data set in a csv file. For example, we have converted a breast cancer relational data base in a csv file.
4. **Data set generation:** This functionality allows the definition of a data set for experimentation. The input is the original data base where the medical information is contained (cases). The output, a set of different data sets that allow training and testing the system according to the experimentation procedure (for example, cross-validation).
5. **CBR application:** runs the CBR experiment according to a given configuration file. As a results, several internal files are generated that contain the outputs of the application and other result-visualization supporting information. All of the files involved in a run are kept internally by the system (with an internal code): configuration file, input file (data sets), output files, and others. So in each moment, the experiment can be replicated, even if

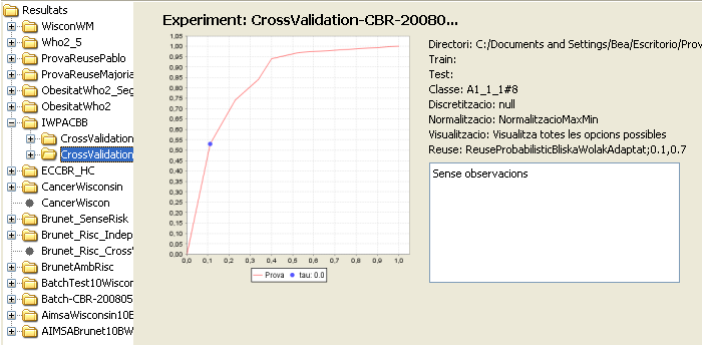


Fig. 2. Experimenter navigator window

the user has removed the files from the directory originally indicated in the configuration file.

6. **Experimenter navigator:** This facility allows the users to navigate in a friendly manner across the different experiments run so far. Figure 2 shows a snapshot of the experimenter navigator. On the left panel, the different experiments executed are listed. When the user moves the cursor across the different experiments, the corresponding results are shown on the top right panel (e.g. ROC curve in Figure 2).

The experimenter navigator is perhaps the most innovative issue in our framework. The explicitly consideration of an independent data functionality is also new.

4 Application to Breast Cancer Diagnosis

Our framework has been developed using the Java language and the jfreechart library [2]. It is compatible with Linux and Windows operating system platforms. The first application we are trying to complete with eXiT*CBR is a breast cancer case-based system.

First of all, we have converted our original access file into a csv file thanks to the data conversion procedure. We have continued by generating different datasets in order to perform a cross validation experimentation. Then, we have set up the CBR system thanks to the edit configuration functionality. The results finally obtained are the ones shown in the navigation window of Figure 2. Figure 3 illustrates the process followed. Unfilled boxes show optional steps, in case we wish to change the destination directory of the results, and the use of another data mining method (plug in).

Next, we realize that the original dataset have dependent data. Therefore, we have run the data independization functionality in order to remove dependent cases and obtaining a new dataset. We run again the experiment with the new data, obtaining new results. Both experiments, have been enlarged by using the corresponding button of the experimenter navigator, and their comparison can be seen in Figure 4. Thus, the tool has facilitated the CBR development and experimentation.

5 Related Work

There are several previous works in which general CBR tools have been developed. For example, cf [3] is a lisp environment designed to be used for rapid prototyping. CBR* [10] and CBR Shell [2] both have been designed according to an object oriented methodology and implemented in java, as our system. Thus, the modularity, reusability and extensionality properties of the object oriented paradigm has been inherited in the CBR frameworks. The main difference of

[2] http://www.jfree.org/jfreechart/

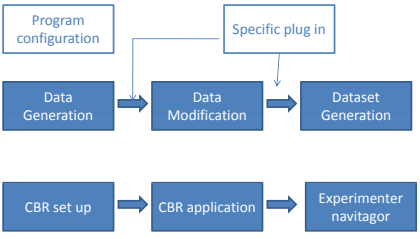


Fig. 3. Main steps followed to develop the breast cancer application

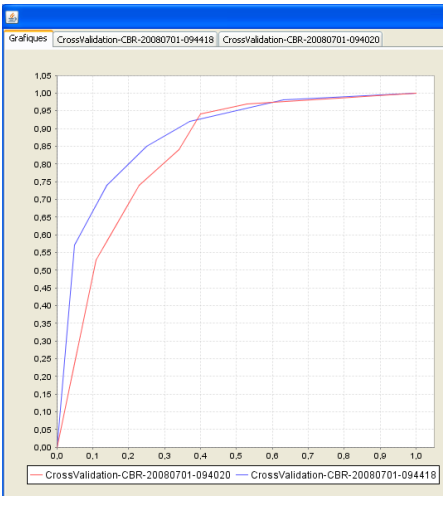


Fig. 4. Comparison of two experiments through the experimenter navigator enlargement facility

the previous approaches is their goal. Our framework is designed to provide to engineers and physicians navigation functionalities across different experiments, helping them to choose the appropriate CBR methods and parameters, and assuring experiment reproducibility. Additionally, we are focussing on a particular case of CBR: medical diagnosis. So we are considering CBR for classification and the visualization techniques are mainly based on ROC graphs, the most commonly used technique in this domain.

6 Conclusions

The development of a CBR system for medical diagnosis purpose is always a time consuming activity. Even that several tools have been developed to facilitate the rapid construction of prototypes, none of them has contemplate the navigation through the experiments of different CBR configurations. The experiment is a

cornerstone in the empirical approach of engineers and physicians to acquiring deeper knowledge about the data they have.

The tool we have developed, eXiT*CBR, tries to fill this gap. We believe that our development and experimental framework would facilitate the interactions with the physicians, without getting lost in the different experiments generated. In addition, our tool allows reproducibility, since we are able to replicate the experiments as many times as required. We think that this is an important issue when dealing with research results concerning medical applications, and we encourage to the medical community to use this kind of tools. We are currently applying eXiT*CBR for developing a breast cancer CBR diagnostic system, as illustrated along the paper.

Acknowledgments

This research project has been partially funded by the Spanish MEC project DPI-2005-08922-CO2-02, Girona Biomedical Research Institute (IdiBGi) project GRCT41 and DURSI AGAUR SGR 00296 (AEDS).

References

1. Aha, D.W., Kibler, D., Albert, M.: Instance based learning algorithms. Machine Learning 6, 37–66 (1991)
2. Aitken, S.: Cbr shell java-v1.0, http://www.aiai.ed.ac.uk/project/cbr/cbrtools.html
3. Arcos, J.L.: cf development framework, http://www.iiia.csic.es/~arcos/
4. Bichindaritz, I., Montinali, S., Portinali, L.: Special issue on case-based reasoning in the health sciences. Applied Intelligence (28), 207–209 (2008)
5. Bilska-Wolak, A.O., Floyd Jr., C.E.: Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with bi-radstm lexicon. Medical Physics 29(9), 2090–2100 (2002)
6. Bilska-Wolak Jr., A.O., Floyd, C.E., Nolte, L.W., Lo, J.Y.: Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning. Medical Physics 30(5), 949–958 (2003)
7. Demar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research (7), 1–30 (2006)
8. Fawcett, T.: Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs (2003)
9. Hammond, K.J.: Explaining and repairing plans that fail. Artif. Intell. 45(1-2), 173–228 (1990)
10. Jaczynski, M.: A framework for the management of past experiences with time-extended situations. In: CIKM, pp. 32–39 (1997)
11. Jurisica, I., Glasgow, J.: Applications of case-based reasoning in molecular biology. AI Magazine 25(1), 85–95 (2004)
12. Lavrac, N., Keravnou, E., Zupan, B.: Intelligent Data Analysis in Medicine and Pharmacology. Kluwer Academic Publishers, Dordrecht (1997)
13. Martinez, T.: Selecció d'atributs i manteniment de la les base de casos per a la diagnosi de falles. Master's thesis, Universitat de Girona (2007)

14. Montani, S.: Exploring new roles for case-based reasoning in heterogeneous ai systems for medical decision support. Appl. Intelligence (28), 275–285 (2008)
15. Francisco Nez, H.F.: Feature Weighting in Plain Case-Based Reasoning. PhD thesis, Technical University of Catalonia (2004)
16. Orengo, C.A., Jones, D.T., Thornton, J.M.: Bioinformatics. Genes, Proteins & Computers. BIOS Sientific Publishers (2004)
17. Perner, P.: Intelligent data analysis in medicine - recent advances. Artificial Intelligence in Medicine (37), 1–5 (2006)
18. Plaza, E., McGinty, L.: Distributed case-based reasoning. The Knowledge Engineering Review 20(3), 261–265 (2005)
19. Pous, C., Pla, A., Gay, P., López, B.: exit*cbr: A framework for case-based medical diagnosis development and experimentation. In: ICDM Workshop on Data Mining in Life Sciences (2008) (accepted, in press)
20. Torra, V., Narukawa, Y.: Modeling Decisions: Information Fusion and Aggregation Operators. Springer, Heidelberg (2007)
21. Wilson, D., Martinez, T.: Reduction techniques for instance-based learning algorithms. Machine Learning 38(3), 257–286 (2000)
22. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6, 1–34 (1997)
23. Zhang, S., Qin, Z., Ling, C.X., Sheng, S.: Missing is useful: Missing values in cost-sensitive decision trees. IEEE Transactions on Knowledge and Data Engineering 17(12), 1689–1693 (2005)

Identification of Relevant Knowledge for Characterizing the Melanoma Domain

Ruben Nicolas[1], Elisabet Golobardes[1], Albert Fornells[1], Susana Puig[2], Cristina Carrera[2], and Josep Malvehy[2]

[1] Grup de Recerca en Sistemes Intel·ligents, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Quatre Camins 2, 08022, Barcelona, Spain

[2] Melanoma Unit, Dermatology Department, Hospital Clinic i Provincial de Barcelona, IDIBAPS, Barcelona, Spain

{rnicolas,elisabet,afornells}@salle.url.edu,
{spuig,ccarrera,jmalvehy}@clinic.ub.es

Abstract. Melanoma is one of the most important cancers to study in our current social context. This kind of cancer has increased its frequency in the last few years and its mortality is around twenty percent if it is not early treated. In order to improve the early diagnosis, the problem characterization using Machine Learning (ML) is crucial to identify melanoma patterns. Therefore we need to organize the data so that we can apply ML on it. This paper presents a detailed characterization based on the most relevant knowledge in melanomas problem and how to relate them to apply Data Mining techniques to aid medical diagnosis in melanoma and improve the research in this field.

Keywords: Health Information Systems, Knowledge Management and Decision Support Systems, Melanoma domain, Computer Aided Systems.

1 Introduction

Melanoma is now one of the most social interesting cancers because it is more frequent in our society and affects people of any age. According to the *American Cancer Society* although is not the most common skin cancer, it is which causes most deaths. This increase, caused by solar habits, makes crucial the early diagnosis, even more if we analyze that this cancer is mortal in approximately twenty percent of cases [1,2] and prompt diagnosis permits practically a secure regain.

Nowadays domain experts works with some plain data bases with disjoined information that does not permit roomy experiments to identify melanoma patterns. In fact, these days diagnosis is not aided for computer systems and requires checking several reports, from different specialists, to give a unified prognostic. Dermatology experts from *Hospital Clinic i Provincial de Barcelona* (HCPB) want to do statistical analysis of its data, and characterize new melanoma patterns.

One way to improve the early diagnosis is to use Knowledge Management and Decision Support Systems based on the statistical results of the cancer information. This imply a Data Mining (DM) problem were we have to analyze high dimensional data, with uncertainty and missing values, to extract the patterns for aid decision making [3]. Some of these techniques have been proved in other kinds of cancer, like the

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 55–59, 2009.
springerlink.com

breast one [4], with results that permits promising investigation. But this technology needs a good data characterization and organization to work and this is non-existent in its domain. To create it we have analyzed data from more than three thousand melanoma patients with information from reports of dermatologists, oncologists, surgeons, pathologists, and other specialists that work at HCPB. Some of these data are used in international studies that analyze specific aspects of the domain. But medical researchers want to study it on the whole to assess if clear patterns are reachable. This aim makes these data really unique and valuable for its variety and completeness.

Then, we have to propose a data organization in melanoma domain to find out which kind of information is relevant and how it is related. This proposal will be the base line for building a DM tool to help experts in melanoma research. Now we want to apply clustering techniques to divide the domain, as is intended previously with less data [5], and use Case-Based Reasoning to support the diagnosis.

The paper is organized as follows. Section 2 analyzes the framework in the field we would like to contribute. Section 3 describes the melanoma domain characterization. Finally, in sections 4 and 5 we discuss about the problem and mark the future work.

2 Related Work

Current works in the melanoma study treat the issue as a sectioned one. We could found works which use only some clinical and pathological data, but with a large number of cases [6], or that studies particular types of information over specific individuals [7,8]. This kind of reference is based on partial databases which are centred on concrete information that does not use the entire data of the domain. A support system that permits us to put it at stake could open new targets in the future using DM. Since early diagnosis of melanoma is the most important factor in the progress of the disease, the diagnostic accuracy is of major interest; is required to establish concise diagnostic criteria based on the clinical information. If we want to apply computer aided systems for getting these objectives, we need a complete and consistent relational model.

3 Melanoma Framework

To apply DM techniques in order to extract patterns from the complete data of the melanoma domain we have to unify the data from different plain databases, from diverse medical specialities, medical information from non electronic support and new data from specific diagnostic machines. This section describes the complexity and constrains of melanoma domain. But, a basic conceptual explanation of the domain is not enough to represent all the relations, specific restrictions, and rules to follow in data insertion, maintenance, and search. Then, with all the collected information we have planed a relational model shown in Fig.1. This model permits all the storage capacities we need, respect the relations and prevent the insertion errors to avoid redundancies. Experts consider that the next aspects explain the domain: 1) person and family, 2) generic medical information, 3) tumors, 4) metastasis, and 5) controls and studies.

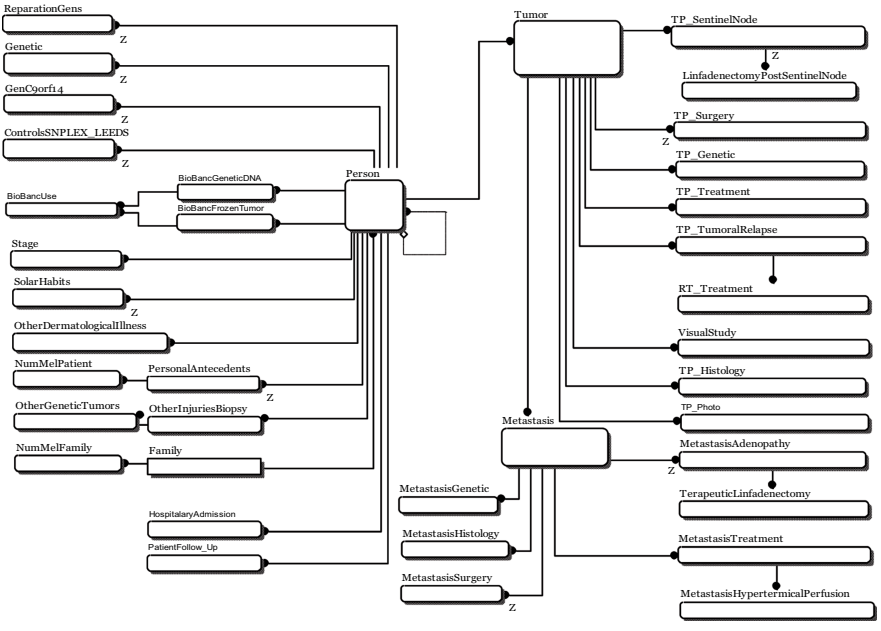


Fig. 1. Relational Model of Melanomas Domain

Person and Family. This aspect summarizes the personal data and antecedents of a person. We note that there are different kinds of person: melanoma patients, familiars of these patients and control cases (healthy people). Patients are also related with their DNA samples and the hospitalary admissions and follow-up that permits to check the illness evolution. Emphasized information about the patient is to know the stage of the illness in different moments (indexed per date), the solar exposure, other dermatological illness, and its biopsies. Another important data to take into account is the familiar relations because it is possible a familiar development of the melanoma.

Tumors. This entity is related with a unique patient, but one patient could have more than one tumor. In addition, this entity has relation with its histologies, treatments applied, genetical information, surgery data, sentinel node, and metastasis. We also have the relation between the tumor and its different images.

Metastasis. This is a concept that depends on a tumor. Each tumor could have different metastasis. The general information is at the same time in relation with histologies, treatments, surgery, genetics, and lymph node metastasis that is included with the addition of linfadenectomy as a particular idea of a metastasis.

Controls and Studies. These aspects refer to the research studies developed by experts in the identification of patterns. A study is the analysis of a certain aspect of a patient. This concept is strongly linked with the person relation because they are who take part in the studies.

4 Discussion

The aim of our current research is to obtain clear melanoma patterns based on all the data existent in this domain. We would like to clusterize the complete information in order to create patterns of diverse characteristics that permit an easy prognostic from different kinds of melanoma even for non-expert doctors. We have also the target of obtaining a good ontological description that allows a better use of CBR techniques to aid medicians in diagnostic accuracy. But all this aims need a previous step that is the characterization of the domain that permits the desired investigation.

The characterisation exposed in section 3 structures melanoma domain and permits not only the target of allowing subsequent work of aid melanomas early diagnosis and accuracy, but it has a suitable insertion of information. This is important because we have found during the process that the previous data have inconsistencies or not related information that come from a bad pick up of it. The incorporation of triggers permits to control this situation and, in addition, filter the existent one in the migration process. Then, at the end of the process, we have obtained a new database completely consistent and with well related information.

We would like to note that the elaboration of the model adds new data obtained from paper reports and diverse databases not unified since now. This situation permits to prepare complete sets of raw data that allows the obtaining of melanoma patterns and statistics for medical research.

5 Conclusions and Further Work

Melanoma is a dermatological cancer with an important impact in our society. Medical researchers in this field want to use techniques to aid its diagnosis. The first step to permit the use of these techniques is the characterization of the domain in order to use the complete data in its studies. The definition proposed permits the creation of an application that allows: 1) An easy data introduction from the specialists, even during attendance work. 2) Knowing the trustworthiness grade of data in order to consider or not low reliable data in certain studies. 3) Rapid creation of new experiments and statistical results in medical research. 4) Study relations between different melanomas information in the bosom of a family. 5) Apply Data Mining Techniques in order to aid specialist to obtain optimal descriptions of patterns to improve the early diagnosis.

Nowadays, our work is focus in the final implementation of the application and the migration of the previous data. After this migration, we could apply Data Mining techniques to create a complete Knowledge Management and Decision Support System *ad hoc* to melanoma's domain that permits to help the medical researchers in the prognostic of this illness. This work are planed in two phases, the first one the extraction of patterns and explanations from this complete data, enlarging previous works that use parts of it [5]; and the second that is the implementation of a hierarchic reasoner that permits to aid the diagnosis by analysing the different kinds of information in separated stages to conclude an unified result.

Acknowledgements

We thank the Spanish Government for the support in MID-CBR project under grant TIN2006-15140-C03 and the Generalitat de Catalunya for the support under grants 2005SGR-302 and 2008FI_B 00499. We thank Enginyeria i Arquitectura La Salle of Ramon Llull University for the support to our research group. The work performed by S. Puig and J. Malvehy is partially supported by: Fondo de Investigaciones Sanitarias (FIS), grant 0019/03 and 06/0265; Network of Excellence, 018702 GenoMel from the CE.

References

1. Gutiérrez, R.M., Cortés, N.: Confronting melanoma in the 21st century. Med. Cutan. Iber. Lat. Am. 35(1), 3–13 (2007)
2. Tucker, M.A., Goldstein, A.M.: Melanoma etiology: where are we? Oncogene 22, 3042–3052 (2003)
3. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2006)
4. Fornells, A., Golobardes, E., Bernadó, E., Martí, J.: Decision Support System for Breast Cancer Diagnosis by a Meta-Learning Approach based on Grammar Evolution. In: 9th ICEIS, pp. 222–227. INSTICC Press (2006)
5. Fornells, A., Armengol, E., Golobardes, E., Puig, S., Malvehy, J.: Experiences Using Clustering and Generalizations for Knowledge Discovery in Melanomas Domain. In: ICDM 2008. 7th Industrial Conference on Data Mining. LNCS, pp. 57–71. Springer, Heidelberg (2008)
6. Balch, C.M., et al.: Prognostic Factors Analysis of 17,600 Melanoma Patients: Validation of the American Joint Committee on Cancer Melanoma Staging System. J. Clin. Oncol.– Am. Soc. Clin. Oncol. 19, 3622–3634 (2001)
7. Puig, S.: Role of the CDKN2A locus in patients with multiple primary melanomas. J. Clin. Oncol. 23(13), 3043–3051 (2005)
8. Malvehy, J., et al.: On behalf of the International Dermoscopy Society Board members. Dermoscopy report: Proposal for standardization Results of a consensus meeting of the International Dermoscopy Society. J. Am. Acad. Dermatol. 57(1), 84–95 (2007)

TAT-NIDS: An Immune-Based Anomaly Detection Architecture for Network Intrusion Detection

Mário Antunes[1] and Manuel Correia[2]

[1] School of Technology and Management - Polytechnic Institute of Leiria, Alto do Vieiro, 2411-901 Leiria, Portugal
mario.antunes@estg.ipleiria.pt
[2] Faculty of Sciences - University of Porto, Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal
mcc@dcc.fc.up.pt

Summary. One emergent, widely used metaphor and rich source of inspiration for computer security has been the vertebrate Immune System (IS). This is mainly due to its intrinsic nature of having to constantly protect the body against harm inflicted by external (*non-self*) harmful entities. The bridge between metaphor and the reality of new practical systems for anomaly detection is cemented by recent biological advancements and new proposed theories on the dynamics of immune cells by the field of theoretical immunology. In this paper we present a work in progress research on the deployment of an immune-inspired architecture, based on Grossman's Tunable Activation Threshold (TAT) hypothesis, for temporal anomaly detection, where there is a strict temporal ordering on the data, such as network intrusion detection. We start by briefly describing the overall architecture. Then, we present some preliminary results obtained in a production network. Finally, we conclude by presenting the main lines of research we intend to pursue in the near future.

Keywords: artificial immune system, tunable activation threshold, network intrusion detection, anomaly detection.

1 Introduction

The vertebrate Immune System (IS) [3] is an appealing metaphor and a very rich source of inspiration for new ideas on anomaly detection applied to network intrusion detection. It possesses two main layers of defense, termed *innate* and *adaptive*, whose main functions are to actively protect the body from intrusions of *pathogens*. The *innate* part of the immune system only recognizes specific known intruders by their "*signatures*", and its behavior is more less the same for all normal individuals in a given species. To overcome this limitation, the *adaptive* immune system has the ability to deal with a much more specific recognition of pathogens and to behave adaptively in order to detect heretofore unseen forms of intrusion. The IS has a complex set of different cell types that interact with each other by the means of chemical messages exchanges. The Antigen Presenting Cell (APC) digests and destroys the microorganisms into small *peptides* and then presents them to *lymphocytes* (*T-cells* and *B-cells*). These cells have specific *receptors* that can *bind* with a certain degree of affinity to the peptides present on the surface of each APC. Depending on the affinity level with the pathogen and on their activation threshold, the cells can become activated, thus initiating an immune response.

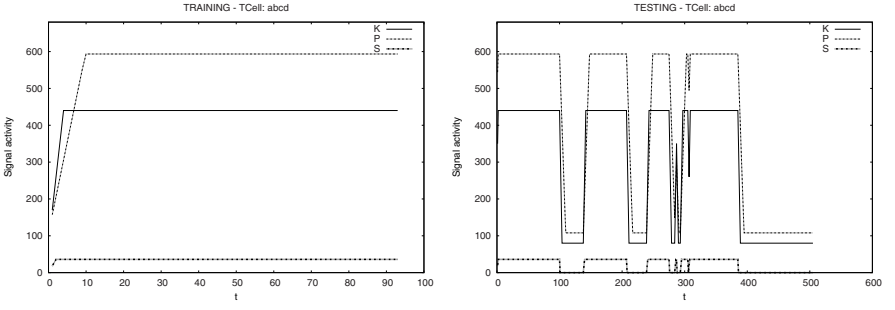
J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 60–67, 2009.
springerlink.com

A Network Intrusion Detection System (NIDS) is an application that monitors a computer network and identifies actions that can compromise its integrity and availability. Its main goal is to positively identify possible occurrences of ongoing attacks and, at the same time, to not be misled by false alarms. NIDS are usually classified into two main types: anomaly, behavior based and misuse, signature based (such as snort-IDS [2]). Metaphorically speaking, the IS challenges are very similar to those faced by a NIDS. Their major goals are to analyze in real time, the iterations of the "system" with the environment and to distinguish what corresponds to legitimate activities (self) from those that manifests themselves as potentially harmful actions (non-self). The adaptive IS behavior can also be seen as an anomaly detector, able to distinguish self from non-self activities. It acquires a "self" profile activity and, during the individual lifetime, adjusts the meaning of self according to the changes occurred in the environment, as well as its level of responsiveness. In the context of network intrusion detection, the meaning of self also changes dynamically through time.

The IS self-non-self discrimination processes inspired the Artificial Immune System (AIS) [6] research community to develop immune-inspired NIDS. There are two main types of immune-based NIDS developed so far [10]: those derived from classical Burnet's Negative Selection (NS) [4] and those that take advantage of Matzinger's Danger Theory (DT) [11]. Forrest's seminal work [7] and Kim's experiments [9] are the most relevant works using NS. The use of DT in the context of NIDS is being developed by Aickelin and colleagues [1]. More recently it has been demonstrated [13,14] these immunological approaches have some serious limitations. The central problem is that there are no prior expectations that they can deal well with change. From one side, NS posits that evolutionary adaptation should have *fixed* immune system cells activation thresholds to an optimal value to ensure efficient self-nonself discrimination. DT bases its activity in built-in *danger* detectors and should eventually fail to detect harmful intrusions that avoid manifesting any of those. These limitations motivated us to investigate the appropriateness of developing an anomaly detection NIDS based on a different view of the IS activity: the Grossman's Tunable Activation Threshold (TAT) hypothesis [8]. The central idea behind TAT is that each immune cell has a tunable activation threshold whose value reflects the recent history of interactions with the surrounding environment. The potentially autoimmune (self) lymphocytes, which are continuously exposed to body antigens, raise their activation threshold and become unresponsive. In contrast, lymphocytes that are not auto reactive and recognize external microorganisms, have low thresholds and are fully responsive upon infection. Thus, the "classification" made by TAT is dynamic and depends only on the present intensity and rate of change of these signals. In summary, TAT requires no prior "classification" of the signals as either "self" or "non-self", and it is expected to automatically adjust each one of the individual cell dynamics into the current environment. The relevance of our research is two-fold. On the one hand, we investigate whether TAT possesses adequate adaptive characteristics to make it suitable to the *non-biological* environment of a computer network. On the other hand, we expect that results obtained within this context will lead to a better understanding of the scope of the TAT hypothesis as a valid and more coherent explanation of the behavior observed in a real vertebrate IS.

2 The TAT Model

We adopted a minimal mathematical model of TAT for T-cells [5]. In this model, T-cell activation is controlled by two enzymes that respond to antigenic signals (S) delivered by the APC: Kinase (K) and Phosphatase (P). The TAT dynamics for a T-cell in two different situations is depicted in Figure 1.



(a) (b)

Fig. 1. TAT dynamics of two individual T-cells in the repertoire

In Figure 1(a) the T-cell receives a constant signal. Except for the initial transient, the condition $P > K$ is fulfilled for the remaining period, meaning that the T-cell will remain inactive. In Figure 1(b) the T-cell receives variable signals and adjusts K and P levels accordingly. During the transient periods in which the signaling is minimal, P and K levels tend towards the initial level, thus decreasing the activation threshold, in such a way that when the signaling increases again, K can transiently supersedes P, thus leading to repeated events of T-cell activation.

The T-cell dynamics exposed by these figures suggests the need to optimize some parameters in order to have good detection accuracy: the *slopes* of K and P (ϕK and ϕP), the maximum values for K and P (K_{max} and P_{max}), the minimum values for K and P (K_0 and P_0) and the affinity threshold between T-cells and peptides.

3 TAT-Based NIDS

The use of IS principles, models and components (such as cells and molecules) in problem solving, is framed into an AIS, whose basic architecture is generally composed by the following basic elements: *representation* of the immune system components, *evaluation* of its interaction within the environment and with each other, and *adaptation* of each component over time [6]. In this section we identify these key elements for an AIS based on TAT and describe the main building blocks for the developed TAT-based NIDS, depicted in Figure 2.

3.1 The Framework

The immune components represented in TAT-NIDS are the APCs, peptides and T-cells. The artificial APC, denoted APC, is a collection of *raw* network packets

("antigens") collected by the network interface (*Network traffic collector* module), encoded in *base64* and then concatenated together in a timely sequence with a predefined duration. This base64 APC stream is spliced into small sized strings ($PEPTIDES$), each corresponding to an artificial peptide. Each artificial cell, termed $TCELL$, is an object that receives signals from the peptides contained in APCs, compares them to a local string representing its unique specificity and adjusts its response threshold by tuning the values of its K and P variables. Table 1 summarizes the metaphor of IS terms in the context of the NIDS architecture proposed.

Table 1. The main immunological terms used in TAT model and its correspondence to our TAT-NIDS system

Immune System	**Network IDS**
Antigen Presenting Cell	APC
Peptide	$PEPTIDE$
MHC/Peptide complex	$PEPTIDE$ Pattern
T-Cell	$TCELL$
T-Cell Receptor (TCR)	$TCELL$ Pattern
T-Cell Repertoire	$TCELL$ Repertoire
Phagocytosis	Pre-processing
Specific recognition	Affinity ($TCELL$ and $PEPTIDE$)
Antigen	Packet
Autoimmune response	False positive
Self	"Normal" network packets
Non-self	possibly an ongoing attack

The core of the architecture is the *TAT-based AIS simulator*. It corresponds to a $TCELL$ dynamics simulator based on the TAT model. The APCs and $TCELL$s interactions and adaptation rate are calculated over time. The simulator processes two kinds of data-sets: a training and a testing data-set. The training data-set is further divided into two parts: a *calibration* and a *validation* part. Both have APCs with normal traffic but the later also has APCs classified as anomalous (Figure 2). In training mode the simulator loads the calibration part and interacts with a non-linear meta-heuristic simplex optimizer [12] that presents the simulator with a set of TAT parameters (Section 2) for minimizing the false alarms rate and simultaneously detecting the attacks included in the validation part. The main reason behind the introduction of attacks in training is to better guide the optimizer in finding a set of parameters that, not only minimizes the rate of false alarms, but can also achieve a low rate for false negatives for the network it has been trained to recognize. Otherwise, the parameters obtained are very likely to be too permissive and inefficient for the testing phase. Finally, in testing mode, the best parameters set obtained are then used to detect the anomalies present in the testing data-set.

The *artificial anomaly generator* module takes advantage of several network tools (such as `nmap`) to generate artificial network packets flows that match snort-IDS [2] rules or are caught by some snort-IDS preprocessing module.

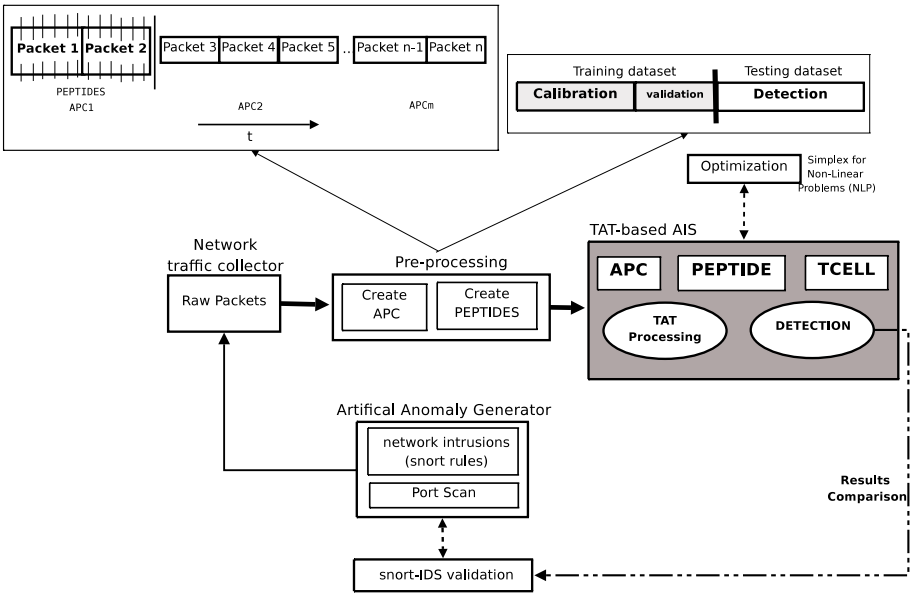


Fig. 2. General architecture of the TAT based NIDS

3.2 The Algorithm

The TAT based detection algorithm is depicted in Figure 3 and can be summarized as follow:

1. APCs are processed sequentially in a temporal basis.
2. Each APC presents its $PEPTIDE$s to the current repertoire of $TCELLS$. If no cell binds the $PEPTIDE$ with a strong affinity, a new one representing the

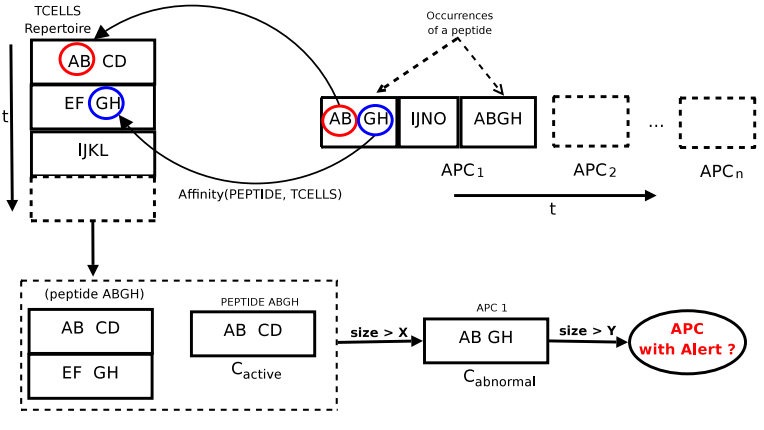


Fig. 3. The steps of the detection algorithm

$PEPTIDE$ is inserted into the repertoire. Otherwise, $TCELL$s that bind with the $PEPTIDE$ are stimulated with a signal (S).
3. According to TAT dynamics, some of those $TCELL$s will become *activated*, with $K > P$ (C_{active}).
4. If the ratio between the number of activated $TCELL$s and the total that bind with a $PEPTIDE$ is higher than a predefined threshold (X), then the $PEPTIDE$ is *classified* as abnormal ($C_{abnormal}$).
5. After processing all the $PEPTIDE$s in an APC, if the number of "abnormal" $PEPTIDE$s is higher than a percentage of the total number of $PEPTIDE$s in the APC (Y), then an alert is raised, indicating that the APC is probably abnormal (non-self).

The classification of an APC is then decided by the "committee" of $TCELL$s that become activated (with $K > P$). This is the role of the committees C_{active} and $C_{abnormal}$, that decide respectively on the classification of a particular peptide as *rare* and the whole of an APC as *abnormal*. These committee *thresholds*, for classifying peptides and APCs, are two parameters also optimized during the training phase.

4 Results

In this section we report preliminary results obtained by TAT-NIDS in two simple network experiments with intranet traffic. For the first we collected about 80 minutes of traffic that we have used for training (approximately 60 minutes for calibration and 20 minutes for validation) and 8 minutes for testing. For the test phase we have generated and inserted *four* attacks. For the simulator parameters optimization (see Section 3) we inserted *one* attack in the training data-set, different from those used for testing.

In the second experience we collected about 60 minutes of traffic that we used for training (approximately 48 minutes for calibration and 12 minutes for validation) and 19 minutes of test. During the test phase we have included an *IP Protocol Port Scan* sweep of the network, made by the *nmap* application. In both experiences, all the artificial attacks generated were detected by the snort-IDS.

With the set of optimized parameters presented in Table 2 we obtained the results described in Table 3. We were able to detect all the attacks (column Quantity) but with some False Positive (FP)s. The TAT algorithm could detect at least one of the attack packet-carrying APCs, which appeared contiguously in the testing data-set. This justifies the different values on column APCs when comparing the true positives and attacks. For instance, in the first experiment the detection of the fours attacks (in which the packets were distributed in 8 APCs), involved only the raising of an alert on 7

Table 2. TAT-NIDS optimized parameter set

Run	K0	P0	S0	Kmax	Pmax	ϕK	ϕP	$Affinity$ %	C_{active} %	$C_{abnormal}$ %
1	80.0	90.6424	8	10.0	11.3303	18	12.545	24.29	53.42	42.35
2	80.0	107.904	8	10.0	13.488	18	9.877	41.56	53.61	28.48

Table 3. Results obtained during the experiments

Run	Phase	APCs	$PEPTIDE$s	$TCELL$s	Attacks		True Positives		False Positives	
					Qty	APCs	Qty	APC	Qty	%
1	Training	916	4,244,899	63	1	2	1	1	5	0.5
	Testing	107	251,472	93	4	8	4	7	6	5.6
2	Training	726	6,387,471	77	1	2	1	1	8	1.1
	Testing	225	419,560	63	2	22	2	6	16	7.1

APCs. Thus, the detection was fully successful, even if not all the APCs were correctly classified.

5 Conclusions

In this paper we have presented an architecture for an AIS-NIDS based on the Grossman's TAT hypothesis. We have also reported some preliminary but promising results, obtained in a real network environment.

We have verified that TAT-NIDS is capable of detecting attack patterns present in the snort-IDS database, without having any prior information about them. These are however very preliminary results and we need to conduct much more extensive experiments with a much larger and richer data-set to be able to measure and better assess the real effectiveness of TAT in network intrusion detection. The major problems we have faced are mainly related to simulator parameter optimization. These happen because the TAT cell simulator parameters space set is extremely large and there is no apparent correlation between them.

We are currently comparing the TAT-NIDS with other AIS for anomaly detection, both in terms of performance and also in terms of accuracy. We are also testing our system on the detection of other vulnerabilities and exploits in the context of much more ambitious data-sets. We are also convinced that the framework we have developed so far can be easily applied, by modifying the preprocessing module that feeds the TAT simulator with APCs, to other complex domains where the objective is to detect some kind of "*temporal*" anomaly, like SPAM classification.

References

1. Aickelin, U., Bentley, P., Cayzer, S., Kim, J., McLeod, J.: Danger theory: The link between ais and ids? In: Timmis, J., Bentley, P.J., Hart, E. (eds.) ICARIS 2003. LNCS, vol. 2787, pp. 147–155. Springer, Heidelberg (2003)
2. Beale, J., Caswell, B.: Snort 2.1 Intrusion Detection. Syngress (2004)
3. Burmester, G.R., Pezzuto, A.: Color Atlas of Immunology. Thieme Medical Publishers (2003)
4. Burnet, F.M.: The Clonal Selection Theory of Acquired Immunity. Vanderbilt University Press (1959)
5. Carneiro, J., Paixão, T., Milutinovic, D., Sousa, J., Leon, K., Gardner, R., Faro, J.: Immunological self-tolerance: Lessons from mathematical modeling. Journal of Computational and Applied Mathematics 184(1), 77–100 (2005)

6. de Castro, L.N., Timmis, J.: Artificial Immune Systems: A New Computational Intelligence Approach. Springer, Heidelberg (2002)
7. Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R.: Self-nonself discrimination in a computer. In: Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy, pp. 201–212 (1994)
8. Grossman, Z., Singer, A.: Tuning of activation thresholds explains flexibility in the selection and development of t cells in the thymus (1996)
9. Kim, J., Bentley, P.: An evaluation of negative selection in an artificial immune system for network intrusion detection. In: Genetic and Evolutionary Computation Conference 2001, pp. 1330–1337 (2001)
10. Kim, J., Bentley, P., Aickelin, U., Greensmith, J., Tedesco, G., Twycross, J.: Immune system approaches to intrusion detection - a review. Natural computing (2007)
11. Matzinger, P.: The Danger Model: A Renewed Sense of Self. Science's STKE 296(5566), 301–305 (2002)
12. Pedroso, J.P.: Simple Metaheuristics Using the Simplex Algorithm for Non-linear Programming. In: Stützle, T., Birattari, M., H. Hoos, H. (eds.) SLS 2007. LNCS, vol. 4638, p. 217. Springer, Heidelberg (2007)
13. Stibor, T., Timmis, J., Eckert, C.: On the appropriateness of negative selection defined over hamming shape-space as a network intrusion detection system. The 2005 IEEE Congress on Evolutionary Computation 2 (2005)
14. Vance, R.E.: Cutting edge commentary: A copernican revolution? doubts about the dangertheory. The Journal of Immunology 165, 1725–1728 (2000)

Novel Computational Methods for Large Scale Genome Comparison

Todd J. Treangen[1,2] and Xavier Messeguer[3]

[1]Institut Pasteur, Microbial Evolutionary Genomics, CNRS, URA2171
[2]UPMC Univ Paris 06, Atelier de BioInformatique, Paris, France
[3]Dept. of Software, Universitat Politècnica de Catalunya, Barcelona, Spain

Summary. The current wealth of available genomic data provides an unprecedented opportunity to compare and contrast evolutionary histories of closely and distantly related organisms. The focus of this dissertation is on developing novel algorithms and software for efficient global and local comparison of multiple genomes and the application of these methods for a biologically relevant case study. The thesis research is organized into three successive phases, specifically: (1) multiple genome alignment of closely related species, (2) local multiple alignment of interspersed repeats, and finally, (3) a comparative genomics case study of *Neisseria*. In Phase 1, we first develop an efficient algorithm and data structure for maximal unique match search in multiple genome sequences. We implement these contributions in an interactive multiple genome comparison and alignment tool, M-GCAT, that can efficiently construct multiple genome comparison frameworks in closely related species. In Phase 2, we present a novel computational method for local multiple alignment of interspersed repeats. Our method for local alignment of interspersed repeats features a novel method for gapped extensions of chained seed matches, joining global multiple alignment with a homology test based on a hidden Markov model (HMM). In Phase 3, using the results from the previous two phases we perform a case study of neisserial genomes by tracking the propagation of repeat sequence elements in attempt to understand why the important pathogens of the neisserial group have sexual exchange of DNA by natural transformation. In conclusion, our global contributions in this dissertation have focused on comparing and contrasting evolutionary histories of related organisms via multiple alignment of genomes.

Keywords: Comparative genomics, genome alignment, interspersed repeats, suffix tree, Hidden Markov Model, DNA uptake sequences, homologous recombination.

1 Introduction

Recent advances in sequence data acquisition technology have provided low-cost sequencing and will continue to fuel the rapid growth of molecular sequence databases. According to the Genomes OnLine Database v2.0 [1], as of June 2008 there are a total of 3825 genome sequencing projects, including 827 that are completed and published. This current and future wealth of available genomic data embodies a wide variety of species, spanning all domains of life . This well documented increase in genome sequence data will allow for unprecedented, in depth studies of evolution in closely related species through multiple whole genome comparisons.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 68–73, 2009.
springerlink.com

Sequence alignment has proven to be a versatile tool for comparing closely and distantly related organisms, nucleotide at a time. However, genome sequences can range in size from millions (i.e. bacteria) up to billions (human genome) of nucleotides, requiring extremely efficient computational methods to perform non-trivial sequence comparisons. Traditionally, sequence alignment methods such as local and global alignment have employed dynamic programming for calculating the optimal alignment between a pair of sequences. While progress has been made [2, 3, 4], optimal global multiple alignment under the sum-of-pairs objective function remains intractable[5] with respect to the number of sequences under comparison, and new heuristics are needed to scale with the dramatic increase of available sequenced genomes.

2 Multiple Genome Alignment of Closely Related Species

Our methodology for multiple genome alignment follows the idea that the comparison of whole genomes can be efficiently based on detecting unique genomic regions, maximal unique matches (MUMs), which appear only once in each genome. We introduce a Suffix Graph data structure that is a modified version of a directed acyclic graph. The most important feature of our Suffix Graph is that it obtains the most compact representation of a suffix tree so far, while maintaining the topology of the structure. We also present an optimized algorithm for finding unique matches (UMs) and maximal unique matches (MUMs) between multiple DNA sequences using a the Suffix Graph data structure. Our approach for searching for unique substrings($\mathcal{UM}$s) yields significant improvements, in both time and space, over existing methods when dealing with multiple DNA sequences. Specifically, given $S_1, \ldots, S_m$ DNA sequences (where S_1 is the smallest genome), we are able to find the $\mathcal{UM}$s among all of the sequence in linear time $O(|S_1| + \cdots + |S_m|)$ and linear space $O(|S_1|)$ Also, the proposed method only requires 8 bytes/character for sequences $S_2, \ldots, S_m$, to retrieve the positions of the $\mathcal{UM}$s in all sequences [6]. Thus this algorithm has linear space complexity with respect to the smallest sequence and linear time complexity with respect to the sum of the length of all sequences. Using these algorithms & data structures, we have designed and implemented an integrated environment for multiple whole genome comparisons based on our MUM search algorithm. The result is **M**ultiple **G**enome **C**omparison and **A**lignment **T**ool, or **M-GCAT** [7]. M-GCAT is able to compare and identify highly conserved regions in up to 20 closely related bacterial species in minutes on a standard computer, and up to 100 in an hour. M-GCAT also incorporates a novel comparative genomics data visualization interface allowing the user to globally and locally examine and inspect the conserved regions and gene annotations (see Figure 1).

3 Local Multiple Alignment of Interspersed Repeats

During the second phase, we shift the focus to *local* multiple alignment. In this phase, we have developed novel computational methods and software for efficient

local multiple sequence alignment of interspersed repeats [8, 9]. Recent advances in sequence data acquisition technology provide low-cost sequencing and will continue to fuel the growth of molecular sequence databases. To cope with advances in data volume, corresponding advances in computational methods are necessary; thus we present an efficient method for local multiple alignment of DNA sequence. Our method for computing local multiple alignments utilizes the MUSCLE multiple alignment algorithm to compute gapped alignments of ungapped multi-match seeds. The method assumes a priori that a fixed number of nucleotides surrounding a seed match are likely to be homologous and, as a result, computes a global multiple alignment on the surrounding region. However, this a priori assumption often proves to be erroneous and results in an alignment of unrelated sequences. In the context of *local* multiple alignment, the fundamental problem with such an approach is that current methods for progressive alignment with iterative refinement compute *global* alignments, i.e. they implicitly assume that input sequences are homologous over their entire length. To resolve the problem, we employ a hidden Markov model able to detect unrelated regions embedded in the global multiple alignment. Unrelated regions are then removed from the alignment and the local-multiple alignment is trimmed to reflect the updated boundaries of homology. We have implemented our method for local multiple alignment of DNA sequence in the `procrastAligner` command-line alignment tool. Experimental results demonstrate that the described method offers a level of alignment accuracy exceeding that of previous methods. Accurately predicting homology boundaries has important implications; for example, tools to build repeat family databases can directly use the alignments without the manual curation required in current approaches and also is likely to aid in the evolutionary analysis of transposon proliferation.

4 Comparative Genomics Case Study of DUS in *Neisseria*

Finally, in this last phase, we employ a comparative genomics approach to study the proliferation of repeat sequence elements in neisserial genomes. Our goal is to understand why the important pathogens of the neisserial group have sexual exchange of DNA by natural transformation. Efficient natural transformation in Neisseria requires the presence of short DNA uptake sequences (DUS), which are highly abundant in their genomes. DUS allow the discrimination between DNA from closely related strains or species and foreign/unrelated DNA. DUS of *Neisseria* spp. is a short signal extending 10 nt, 5'-GCCGTCTGAA-3' [11]. It is present in approximately 2000 copies occupying 1% of the sequenced neisserial genomes. DNA uptake sequences could have proliferated either by selection for transformation or by selfish molecular drive, and through their study we hope to enlighten their evolutionary role. We took advantage of the opportunity provided by the availability of six complete neisserial genomes to globally align the core genome and to define the sets of genes that are ubiquitous and those that were

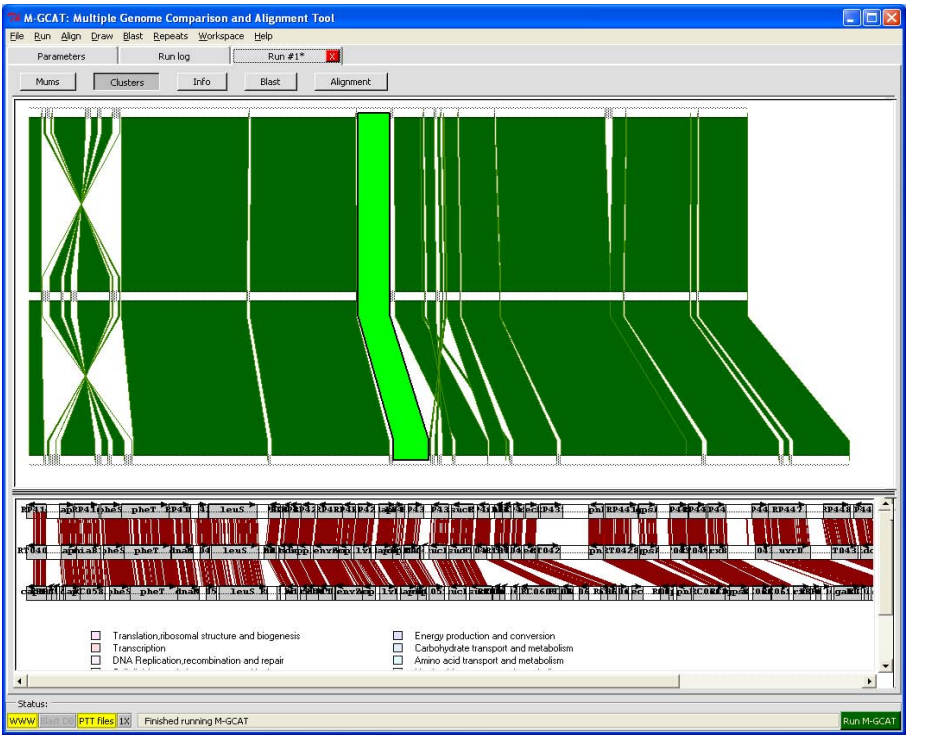


Fig. 1. Visual representation of a multiple genome comparison of 3 *Rickettsia* genomes inside of the M-GCAT interactive genome comparison viewer. There are 22 highly conserved collinear regions displayed, covering approximately 91% of the total genomic sequence. The region highlighted in green and indicated with the black arrow is one of the 22 regions found to be highly conserved among the 3 closely related species. Directly below the global genome comparison viewer, all annotated genes are displayed horizontally as rectangles. The red vertical lines represent the multiple maximal unique matches common to all 3 genomes.

recently acquired or recently lost in each group. Multiple genome alignments provided a new approach in solving the puzzle of the origin and fate of DUS in these genomes, which helped to elucidate the association between these signals and recombination events [10]. A strong correlation between the average distance between DUS and the length of conversion fragments was found, indicating that the process of transformation is tightly linked to and even shaped by a history of recombination.

5 Conclusion

Due to the great amount of DNA sequences currently available, tools which can efficiently and accurately align and compare multiple genomes are essential

for identifying evolutionary patterns. We have proposed novel data structures, algorithms, and software for active areas of research in comparative genomics. Additionally, we have performed an comparative genomics analysis to study the evolutionary role of DNA uptake sequences in six strains of *Neisseria*. The novel algorithmic ideas presented in this doctoral thesis for multiple genome comparison based on global and local alignment allow for multiple genome comparison of organisms at varying evolutionary distances. Our global contributions in this dissertation have focused on comparing and contrasting evolutionary histories of related organisms via their genomes.

Acknowledgments

T.J. Treangen would like to especially thank Eduardo PC Rocha of the Institut Pasteur for his guidance and support throughout the thesis. We would also like to thank Aaron E. Darling, Ole Herman Ambur, Mark A. Ragan, Nicole T. Perna, and Tone Tonjum for their collaboration throughout various phases of the thesis. This work has been supported by the LogicTools project (TIN2004-03382) funded by the Spanish Ministry of Science and Technology and the EU program FEDER and AGAUR Training Grant FI-IQUC-2005.

References

1. Liolos, K., Tavernarakis, N., Hugenholtz, P., Kyrpides, N.: The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Research 34, 332–334 (2006)
2. Edgar, R.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (2004)
3. Thompson, J.D., Higgins, D.G., Gibson, T.: Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680 (1994)
4. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302, 205–217 (2000)
5. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348 (1994)
6. Treangen, T.J., Roset, R., Messeguer, X.: Optimized search for common unique substrings, on both forward and reverse strands, in multiple DNA sequences. In: Poster proceedings of the 1st internation conference on Bioinformatics Research and Development BIRD (2007)
7. Treangen, T.J., Messeguer, X.: M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. BMC Bioinformatics 7, 433 (2006)
8. Darling, A.E., Treangen, T.J., Zhang, L., Kuiken, C., Messeguer, X., Perna, N.T.: Procrastination leads to efficient filtration for local multiple alignment. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175. Springer, Heidelberg (2006)

9. Treangen, T.J., Darling, A.E., Ragan, M.A., Messeguer, X.: Gapped Extension for Local Multiple Alignment of Interspersed DNA Repeats. In: LNBI proceedings of the International Symposium on Bioinformatics Research and Applications ISBRA (2008)
10. Treangen, T.J., Ambur, O.H., Tonjum, T., Rocha, E.P.C.: The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biology 9(3), R60 (2008)
11. Goodman, S.D., Scocca, J.J.: Factors influencing the specific interaction of Neisseria gonorrhoeae with transforming DNA. J. Bacteriol. 173, 5921–5923 (1991)

Improving Literature Searches in Gene Expression Studies

Joel P. Arrais, João G. L.M. Rodrigues, and José Luis Oliveira

University of Aveiro, Campus Universitário de Santiago,
3810-193 Aveiro, Portugal
{jpa,jlo}@ua.pt

Abstract. MEDLINE is the premier literature database in the biomedical and life sciences fields, containing over 17 million references to journal articles. Searching in this database can be performed through PubMed, a web interface designed to provide a rapid and comprehensive retrieval of articles matching a specific criteria. However, considering the complexity of biological systems and of the genotype to phenotype relations, the results retrieved from PubMed can be only a short view of the relevant information that is available. In this paper we present a new approach for expanding the terminology used in each query to enrich the set of documents that are retrieved. We have developed a paper prioritization methodology that, for a given list of genes, expands the search in several biological domains using a mesh of co-related terms, extracts the most relevant results from the literature, and organize them according to domain weighted factors.

Keywords: DNA microarrays, PubMed, query expansion, information retrieval.

1 Introduction

Mainly due to the rapid progress of biotechnology, an increasing number of scientific articles are published in a wide range of life science scientific journals. Despite the benefits of this offspring of data this can also create big issues to any researcher that wants to navigate through the literature and extract information relevant to their study [1-4]. To help in this task, search engines such as the Entrez PubMed (*http://www.pubmed.com*) and the Crossref Search [5] have been proposed. These tools index the major scientific journals allowing users to use a single interface rather than have to search along all the journals' sites.

Despite the major advantages of these tools they do not solve all possible user questions. One of those questions consists in retrieving from the literature, information related to a set of genes which is a common outcome of many experimental techniques. A common example are the gene expression studies, where, after measuring the relative mRNA expression levels of thousands of genes one usually obtain, a subset of differentially expressed genes that are then considered for further analysis [6]. In this case, the ability to rapidly extract from the literature the existent relations among this list of differentially expressed genes is critical, because, despite the big efforts to develop gene expression databases, such as the ArrayExpress [7] and the GEO (Gene Expression Omnibus) [8], the literature is still the main source of information. As so, the development of tools that crawl the existing literature, searching for

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 74–82, 2009.
springerlink.com

relevant information is of utmost importance. One previous approach to address this need has been applied in GeneBrowser [9]. It consisted in selecting all the articles that refer to a specific gene and then ranking the articles according to the affinity to the list of genes that they refer to. Despite the added value of the results, this tool still requires improvements in the selection of the most relevant papers.

In this paper we present a new methodology to select and rank the articles that have strongest relations with the given set of genes and with the aims of the study. This technique consists in using a well defined network of biological concepts to expand the given search criteria - each element from the gene list - to associated concepts in order to expand the queries and to increase the relevance of the results.

2 Extracting Knowledge from DNA Microarray Data

Despite the benefits of the existent data sources and tools to extract knowledge from a set of genes many relevant information are still enclosed in the literature in the form of raw text. This is mainly because many biological databases, that have the capability to store data in a structured form that can be readable and processed by computers, are still filled by manually searching the literature. This way, their updating rate is typically slower than the number of biological evidences that are daily published. Although this process is changing as publishers start to oblige submitters to send the results to well know databases (such as ArrayExpress, for gene expression studies) there is still a gap between the information contained in such databases, and the one stored in the literature, making this last an invaluable source of up-to-date information.

In the field of biomedical and life sciences, the MEDLINE database, of the United States National Library of Medicine, is the most relevant source of information. Its broad collection includes more than 17 million abstracts collected from about 5000 journals dating back from 1960s (*http://www.nlm.nih.gov/pubs/ factsheets/MEDLINE.html*). Every day, roughly 3000 new abstracts are added. To search and retrieve information from this database an easy to use web interface, named PubMed (*http://www.pubmed.com*), has been developed.

The main problem with the MEDLINE/PubMed, is because the information is stored in an unsorted and unstructured form and extracting relevant information is still a challenge. To address that in the past years many information extraction and text mining techniques have been applied (Fig.1). Examples are the AliBaba system [10] that enables to see the PubMed database as a graph, BioProber [11] that uses text mining techniques to enable the discovery of abstracts related with an given list of abstracts, GoPubmed [12] that enables to sort and view the PubMed abstracts according to Gene Ontology categories, PubMed interact [13] that proposes a new interface to put Boolean queries into the PubMed database and iHOP [14], that offers access to the underlying literature by means of a network of genes and proteins.

In the field of gene expression studies some software solutions have already been proposed. For instance, micro-GENIE [15] is a tool that for a set of genes combines information from the UniGene and SWISS-PROT databases to perform semi-automated querying over the PubMed database. This tool has as main drawbacks the limited scope of the used methodology and the long announced time required to perform queries (30 minutes for a set of 200 genes). One last tool worth mentioning,

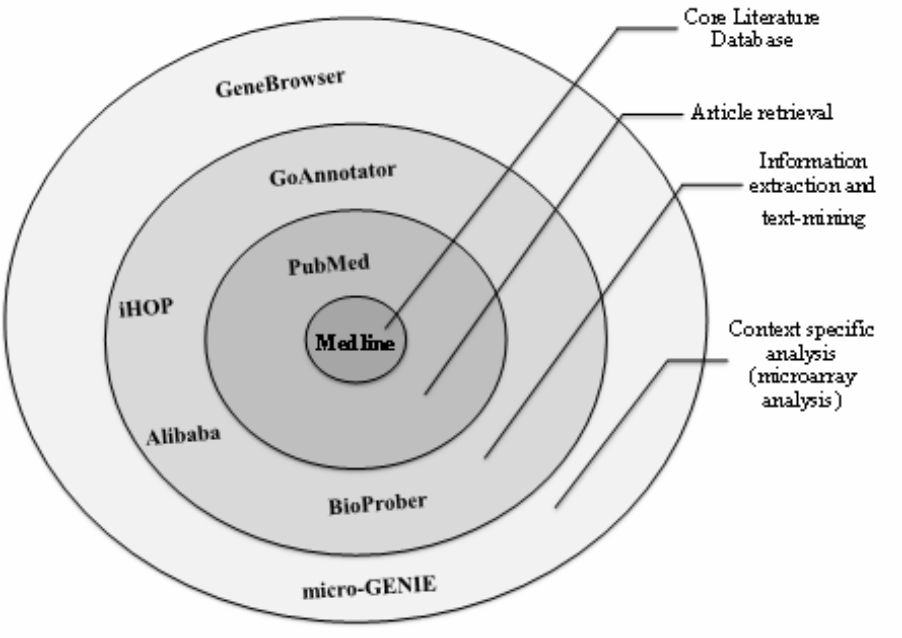


Fig. 1. An overview of the cascade of tools associated with Medline

GeneBrowser, consists in comprehensive web based application that integrates both public biological data and functional analysis tools to study a set of differently expressed genes. One of the features of this tool consists in collecting and showing, in real time, the set of abstracts that are more related with the initial given list of genes. Despite its added value, this system lacks of a more robust approach to select and sort the most relevant abstracts.

The aim of this work is to present an improved methodology to select and rank the most relevant abstracts to a list of genes. The proposed methodology can be compared to an assembly of the tasks performed by researchers when they manually treat and process the data. Indeed, a researcher searching for evidences of the relationship among a list of genes, will often start by inserting this list, or some groups from this list of genes, into the PubMed interface. The previous query will return a huge list of abstracts where all, or some, of the genes are refereed. Simultaneously, tools similar to the FatiGO+ [16] are used to detect the most relevant functional categories within the initial list of genes. These functional categories are then used as search criteria on the original abstract list, in order to refine the search. Iteratively performing this process will lead to a narrowed list of abstracts, likely to contain those of the most interest, however, it will also lead to a large amount of time spent doing so. This time-consuming methodology can be surpassed using computer methods, which was our motivation to develop the proposed tool.

3 Proposed Workflow

Our methodology follows a four-step workflow that begins with a list of genes (Fig. 2a) and ends with a list of PubMed Identifiers (PMIDs) returned to the user (Fig. 2e). The first step is the acceptance and validation of the original gene identifiers, which are mapped to a chosen consistent naming convention: Entrez Gene Identifier. Next, we use query expansion techniques [17, 18] to broaden the list of terms to be later queried (Fig. 2b). This step is made using a previously built network of biological concepts, such as: pathways, proteins and gene synonyms. The list of keywords is then submitted to a search engine which iterates over millions of PubMed records, locally stored on our server, matching each positive hit with the abstract's PMID (Fig. 2c). The last step is the assembly, ranking (Fig. 2d) and displaying of a final list of PMIDs (Fig. 2e).

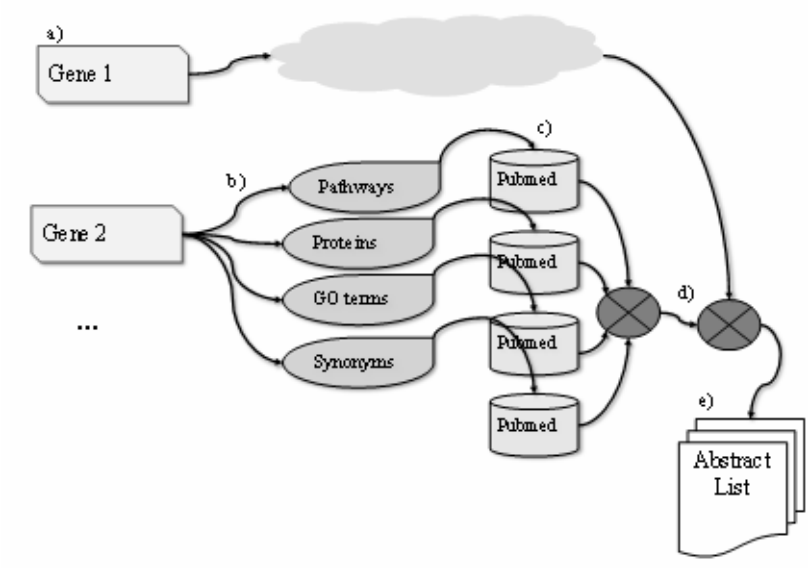


Fig. 2. Proposed workflow: from the genes to the abstract list **a)** Initial list of genes identifiers **b)** Query expansion **c)** Search over the local index of PubMed **d)** Assembly and ranking of the results **e)** Final abstract list returned

3.1 Acquiring, Mapping and Validating Genes Identifiers

The first step of the workflow consists in obtaining and validating the dataset by removing all the terms that did not correspond to a valid gene and mapping the identifiers to a consistent naming convention (numerical identifier from the Entrez Gene database). This mapping step is required because due to the lack of a universal coding schema for genes, each database developed its own naming convention. The solution to overcome the problems caused by this multiplicity of identifiers, consisted in the development of a local database that stores the mapping among several identifiers.

3.2 Using Query Expansion Techniques

Query expansion is commonly defined as the process of reformulating a seed query to improve performance in information retrieval operations [17, 18]. As such, there are plenty of techniques to perform this process, each with a particular approach and specific advantages. Amongst the most usual, there is the expansion of a search string based on synonyms, the search of morphologically-related words and the re-weighting of the terms in the search string. The main motivation to use query expansion is to increase the quality of the search results.

Behind every information retrieval technique, there are two parameters that define its efficiency: recall and precision. Recall measures the fraction of documents that are relevant for the retrieved query when compared to the entire database. By other side, precision measures the fraction of documents that are relevant for the retrieved query when compared to the returned dataset. The ideal situation is to simultaneously have higher values of recall and precision. Usually with the use of query expansion techniques the recall is increased at the expense of precision. One way balancing this tradeoff is with the choice of terms that are conceptually as close as possible with the initial seed term. To address this issue, we choose to use biological concepts to expand the query. The developed system expands the initial query by a simple matching of the initial gene name to the appropriate proteins, pathways and other biologically relevant fields that will be used to formulate query string.

The example in Fig. 3 illustrates this procedure. For the gene YCR005C (Fig. 3a) from *Saccharomyces cerevisiae,* we first obtain its Entrez gene identifier (Fig. 3b) and then we expand for the following terms: gene name (*CIT2*) (Fig. 3c), pathway (sce00020, sce00630) (Fig. 3d), protein (P08679) (Fig. 3e) and the GO (0006101, 0006537, 0006097, etc) (Fig. 3f).

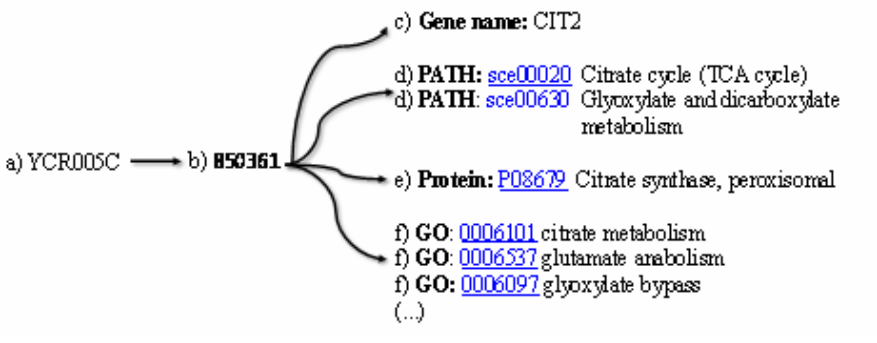


Fig. 3. Query expansion example **a)** Initial gene identifier **b)** Entrez gene identifier **c)** Gene name **d)** Metabolic pathway **e)** Protein **f)** Gene Ontology term

3.3 Searching over PubMed

As with every other remote database, there is a downside in remotely querying PubMed: speed. As there are currently over 17 million citations, it is understandable that accessing in real-time to such a large database would take a long period of time. In

addition, due to the use of query expansion each set of genes will lead, not to one, but to a huge number of effective queries into PubMed.

To overcome this problem we chose to create a local index of the PubMed database, thus making the limiting factor in the search speed the hardware capacity when performing iterations over the local index.

The index was built considering that only the abstract and the title contain relevant information for the search. In addition, in order to identify each abstract in the index, we also extract its PMID. Therefore, and for optimization reasons, our indexing mechanism processes both the title and the abstract in one single field, which is indexed and tokenized, but not stored, while the PMID is stored as an integer, in order to be later retrieved with each positive match in a search.

The search mechanism used over this index is provided by the same software as the indexing mechanism, giving us, a unified indexing-searching system. Each query is constructed in real-time, using the results from the query expansion step, and submitted to the search function, yielding the PMIDs that positively matched the search terms.

3.4 Assembling the Results

Searching for a given set of terms in our local index retrieves a list of PMIDs. Since we are using query expansion techniques to reformulate the query, for each gene, several searches are actually performed, which means that, at the end of our search, we have several lists. These lists need to be assembled in order to produce a final sorted list that is returned to the user. This assembly required a ranking algorithm that should reflect the following two requirements. The first is that the system has to allow the user to define weights for each set of terms originated by each branch of the query expansion (e.g., protein name, pathway name). The other, that the final list of results should reflect the relevance of the articles, so that the most relevant articles come near the top of the list. The formula that reflects the previous requirements is defined as:

$$Rank_i = \sum_{j=1}^{Ntc} W_j \times n_{ij} \quad (1)$$

i specific abstract; *j* specific expansion term class; N_{tc} total number of term classes; W_j weight attributed to the term class *j;* n_{ij} number of times that the expansion to the class *j* produce positive hits in the abstract *i*.

4 System Implementation and Avaibility

The proposed tool has been fully developed with the use of free, open-source technologies. The Phyton programming language (*http://www.python.org*) was chosen to develop the application, while the MySQL database management system (*http://www.mysql.com*) was used to locally store some of the data, namely, the tables containing the query expansion terms. The main motivation to use Python was its good balance between ease of use, number of available libraries and good performance. MySQL was chosen because it is a popular and robust open-source database

management system. In addition it is easy to install and maintain, while offering seamless integration in Phyton scripts as well as good performance.

As we already mentioned, we opted for having a local index of the MEDLINE database, due to performance issues. The indexing was produced with Lucene (v 2.3.0 - *http://lucene.apache.org/*), an open-source information retrieval library, through its port PyLucene. Since no local storage of the abstracts is done, those are obtained in real time, in batches from the PubMed database. This data, in XML, is parsed in order to filter the contents (title and abstract) that are indexed and tokenized. This allows us to gather a significant portion of the whole PubMed database in less than 2GB of hard disk memory.

The construction of the local database, needed for the query expansion, was built through the access to several data-sources. The access to the KEGG and the Gene Ontology databases was performed by the use of webservices to directly populate the database. From the Uniprot/TrEMBL and the Entrez Gene databases a dump files, in raw text format, was downloaded, parsed and the information of interest was stored in the local database.

The searching mechanism also uses Lucene's capabilities. Having only one stored field, specifically an integer, brought significant performance enhancements when compared to previous scenarios, bringing us closer to our objective of a near-instantaneous results list.

The index, the query expansion database as well as all the Python scripts required to test the proposed methodology are freely available at http://bioinformatics.ua.pt/tools/qetool.zip. We kindly invite users to download it, to perform their own searches and to provide feedback and suggestions for improving its features. In addition, other developers can take advantage of the Lucene index and the Python scripts, as a framework to implement their own search mechanisms.

5 Discussion

A wide range of web-based text mining tools offer optimized or, specialized versions of the search tools provided by the PubMed database. Each of the available web based systems have unique features, and may be preferred for particular users or to address specific problems. Despite that, the tool presented here offers, at least, three unique features that, to our knowledge, are not found in any other tool. Together, these characteristics make this a unique tool for addressing the issue of selecting the most relevant papers from a set of genes.

First, the input is a list of genes rather than a simple query string. This characteristic makes the system especially useful to be used in the context of gene expression studies where one, or more, list of genes are usually produced as outcome of an experiment. It is also assumed the existence of relations among the list of genes, being the system responsible for finding the papers that evidence those relations.

Second, the possibility to automatically expand the initial query from a list of genes to a set of other biologically related concepts.

Third, by allowing users to attribute different weights to the query expansion terms the tool allows users to look to the literature from different perspectives. For instance, users interested in studying metabolic pathways can set a maximum weight to this

class. As a result they will obtain the list of papers that have references to the metabolic pathways where the initial list of genes appears.

6 Conclusion

In this paper, we propose a new methodology to extract relevant articles from the available literature on MEDLINE, regarding an initial set of genes. In order to achieve this goal, we combined query expansion techniques with both document indexing and open source software, such as Lucene and MySQL. The results were an overall increase on the relevance of the returned articles. Also of remark is that only open-source software was used to develop this application.

Further development is focused on the creation of an user-friendly web-interface, so that the methodology here presented can better exploited by the scientific community. We also intend to conduct further evaluation and comparison of this methodology with other tools. Nevertheless, we believe the preliminary results seam very promising and surely reflect a major advance when compared with the available methods.

Acknowledgments. J. Arrais is funded by FCT grant SFRH/BD/23837/2005. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2008-2011) the ALERT project.

References

1. Jensen, L.J., Saric, J., Bork, P.: Literature mining for the biologist: from information retrieval to biological discovery. Nat. Rev. Genet. 7, 119–129 (2006)
2. Weeber, M., Kors, J.A., Mons, B.: Online tools to support literature-based discovery in the life sciences. Brief Bioinform. 6, 277–286 (2005)
3. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Brief Bioinform. 6, 57–71 (2005)
4. Krallinger, M., Valencia, A.: Text-mining and information-retrieval services for molecular biology. Genome Biol. 6, 224 (2005)
5. Renn, O.: DOI, CrossRef, LINK, and the future of scientific publishing. J. Orofac. Orthop. 62, 408–409 (2001)
6. Dopazo, J.: Functional interpretation of microarray experiments. Omics 10, 398–410 (2006)
7. Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., Brazma, A.: ArrayExpress–a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 35, D747–D750 (2007)
8. Barrett, T., Edgar, R.: Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 411, 352–369 (2006)
9. Arrais, J., Santos, B., Fernandes, J., Carreto, L., Santos, M.A.S., Oliveira, J.L.: GeneBrowser: an approach for integration and functional classification of genomic data. Journal of Integrative Bioinformatics 4(3) (2007)

10. Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., Leser, U.: AliBaba: PubMed as a graph. Bioinformatics 22, 2444–2445 (2006)
11. Jang, H., Lim, J., Lim, J.H., Park, S.J., Lee, K.C.: BioProber: software system for biomedical relation discovery from PubMed. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., vol. 1, pp. 5779–5782 (2006)
12. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 33, W783–W786 (2005)
13. Muin, M., Fontelo, P.: Technical development of PubMed interact: an improved interface for MEDLINE/PubMed searches. BMC Med. Inform. Decis. Mak. 6, 36 (2006)
14. Hoffmann, R., Valencia, A.: Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21(suppl. 2), ii252–ii258 (2005)
15. Korotkiy, M., Middelburg, R., Dekker, H., van Harmelen, F., Lankelma, J.: A tool for gene expression based PubMed search through combining data sources. Bioinformatics 20, 1980–1982 (2004)
16. Al-Shahrour, F., Minguez, P., Tarraga, J., Medina, I., Alloza, E., Montaner, D., Dopazo, J.: FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Res. (2007)
17. Yonggang, Q., Hans-Peter, F.: Concept based query expansion. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Pittsburgh, Pennsylvania, United States (1993)
18. Jinxi, X., Croft, W.B.: Query expansion using local and global document analysis. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Zurich (1996)

Implementing an Interactive Web-Based DAS Client

Bernat Gel and Xavier Messeguer

Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya (UPC)

Summary. The Distributed Annotation System (DAS) allows clients to access many disperse genome and protein annotation sources in a coordinate manner. Here we present DASGenExp, a web based DAS client for interactive visualisation and exploration of genome based annotations inspired in the Google Maps user interface. The client is easy to use and intuitive and integrates some unique functions not found in other DAS clients: interactivity, multiple genomes at the same time, arbitrary zoom windows,... DASGenExp can be freely accessed at http://gralggen.lsi.upc.edu/recerca/DASgenexp/.

Keywords: DAS, visualisation, genome, browser.

1 Introduction

In recent years, genomic data comprising genome sequences and its annotations have been growing at fast pace. Every year more organisms are sequenced and studied and more data is generated. It is important to note that most of this data is completely free and available on the internet.

As more research groups publish their data and a greater number of specific databases appear it becomes more difficult to work with the available data and even find it. Some of the biggest databases offer data visualisation tools for the user to explore its content but those visualisation systems usually fail to integrate data from different sources. Additionally small groups are setting up project related databases which are hardly maintained and so updates in the genomic assemblies can draw data from two different sources incompatible.

In an attempt to address those issues the Distributed Annotation System (DAS) was proposed. By adding a simple interface to the already existing databases a new abstraction layer was created which converted the diverse and different databases into a unique global distributed database.

Here we present DASGenExp, a web-based visualisation tool designed to interactively explore the bast amount of genomic data available in this emerging distributed database.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 83–91, 2009.
springerlink.com

2 The Distributed Annotation System

The Distributed Annotation System (DAS) [1] was developed to provide an integrated interface to the annotation data from diverse sources. It allows clients to access DAS data sources from different institutions and geographically dispersed using a simple standard protocol and to fetch coherent data from them.

The DAS protocol is based on HTTP and XML and very network friendly. To request information, the client makes an HTTP request encoding all needed parameters in the URL and gets an XML file with the needed data. There are few different URLs to encode different types of requests and the parameters are very consistent between them. This simplicity makes it easy to implement in different languages and, actually, client libraries exist at least for Perl and Java. DAS servers, or sources as they are usually referred, are somewhat more complex but free Perl and Java implementations are also available.

A DAS source must be referenced to a DAS Reference Server so it's positional annotations, the main class of available annotations, are correctly referenced. This is even more important since different versions of the reference sequences exist due to updated reference genomic assembly or new advances in the genome sequenciation. A DAS Reference Server is a special DAS source able to answer to some additional requests, mainly the `sequence` command which is used to retrieve the reference sequence itself. The reference sequences in a DAS source are known as entry points and usually correspond to chromosomes – for genome based sources – and to proteins – in protein based servers –.

There are more than four hundred DAS servers currently active, some of them supported by big institutions as EBI. Additionally, a central repository have been proposed [4] where DAS sources can be registered. Those registered sources can be browsed or searched by clients. Different freely available DAS client have been developed either desktop or web-based. Clients are usually specialised either in genome or protein visualisation. In this section we will refer only to genome based clients.

Both UCSC [2] and Ensembl [3] have added DAS support to their existing web-based genomic browsers. However, DAS integration is somewhat difficult and may not be completely clear how data coming from DAS sources can be added to the browsers. In addition, both browsers suffer from lack of interactivity since the whole representation have to be reloaded in order to move the viewer to a different part of the sequence. Those browsers, however, are very powerful and a reference in genomic browsing nowadays.

3 DASGenExp

DASGenExp is a web-based genome oriented DAS client. It offers an easy to use user interface partially inspired by Google Maps usability and provides some new functionality and options not found in other clients along with a high level of interactivity.

The viewer can integrate genome oriented annotations from any DAS server and draw them in a graphical representation of the genome along with the

reference sequence. This graphical representation can be moved by dragging it with the mouse and zoomed in and out. Zoom can be changed both via a slider or the mouse wheel and ranges from the closest base pair view to a chromosome wide one. Extended information about a feature pops up when the mouse is over it. Double clicking on any point of the genome a new window is created viewing the same region but centred at the clicked base and with th zoom to its closest level.

More than one window, or track container, can be present on the page at the same time. It is possible, for example, to have different zoom levels of the same region synchronized at the same time. It is possible, too, to have two viewer windows displaying two different independent regions from the same or different genomes.

Short Example

A simple example session with DASGenExp could be like that. First of all, open the page `http://gralggen.lsi.upc.edu/recerca/DASgenexp/` in the browser (preferably Firefox). When the page is completely loaded, select 'Homo sapiens - NCBI36' in the first combobox in order to explore the human genome based on the NCBI assembly release 36. A request is then made to get the chromosomes for the selected organism.

Select a chromosome from the second combobox, for example '2', and the sequence of the chromosome will appear. The `Tracks` menu will be enabled. Add two tracks from the karyotype by selecting `Tracks` → `ensembl NCBI36 karyotype` → `band:gneg` and `Tracks` → `ensembl NCBI36 karyotype` → `band:gpos50`. The data for those two tracks will be loaded and a blue box will appear. Drag the view slightly to the right with the mouse to reveal the beginning of the blue box where the name of the feature can be seen: `p25.3`.

Zoom out using the mouse scroll wheel until most of the chromosome can be seen. Drag the image to center it around the '100Mb' tick and then zoom in using either the mouse or the slider on the right up to half the maximum zoom. Then add another track selecting `Tracks` → `ensembl NCBI36 transcript` → `exon:coding`.

Add a zoomed-out view using `Window` → `Window Zoom`. A new window will appear with the same data. If we drag the main window, the linked zoom window will move accordingly.

The final state of the application will be similar to Figure 1.

4 Implementation

DASGenExp is a client-server application with the client being a javascript web application and the server a collection of Perl CGI scripts.

The client runs inside the web browser which acts as the application container. It controls data representation and user interaction. The communication between the client and the server is done via AJAX technology but using JSON

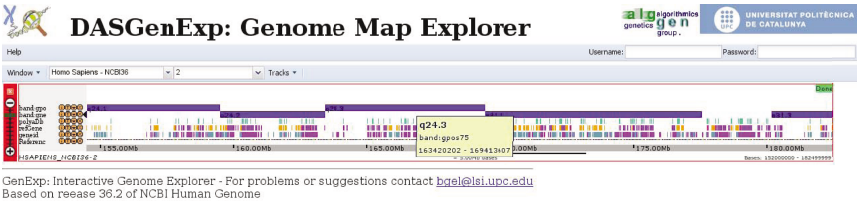


Fig. 1. Screenshot of DASGenExp with some data added

as transport instead of XML as classical AJAX. JSON, JavaScript Object Notation, is somewhat less expressive than XML but its much lightweight and less verbose and so it's very well suited for scenarios where large amounts of very simple data have to be transmitted.

The server subsystem is the one responsible of fetching the actual data from thediversee DAS sources, process it and send it back on clients request. Communication between the server and the sources is made using the DAS protocol (Figure 2).

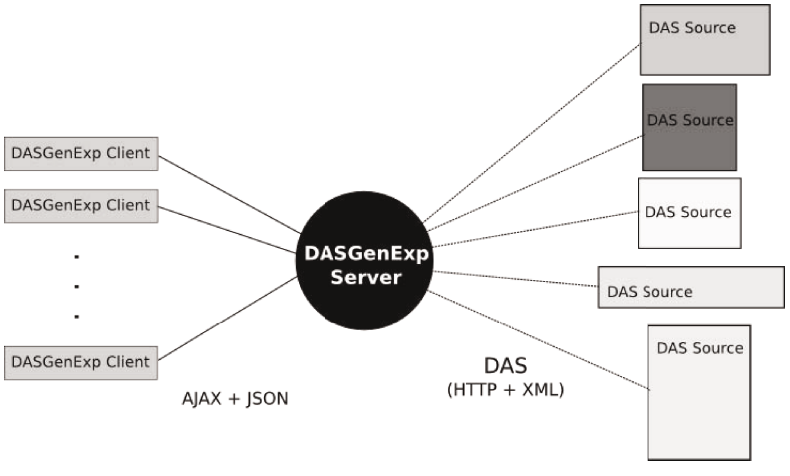


Fig. 2. Scheme of DASGenexp system

4.1 The Client

The most important and complex part of DASGenExp is the web-based client. Entirely written in javascript, the client runs inside the internet browser and is the main responsible of the data visualisation and interaction. It uses AJAX technology to asynchronously request information from the server without ever recharging the entire webpage. In fact, the DASGenExp application is made of only one and fairly simple HTML file coupled with a collection of JavaScript files containing all the application logic.

The client uses two javascript libraries. The Prototype library [5] is used as a base javascript library providing useful and commonly used functionalities such as DOM manipulation and AJAX handling. It also provides shorthands for some often used functions resulting not only in faster and less error-prone programming but also in shorter and more legible code. Another very useful functionality Prototype offers since it's last version is a complete system to manage custom made events in exactly the same way the default events are handled. This event system is very powerful and gives the opportunity to use event based communication between different parts of the system which is very convenient in a GUI based application and even more when there are asynchronous processes involved such as AJAX requests.

The second library used, EXT [6], offers a full range of very interactive and desktop-looking widgets and a modular object-oriented framework to handle them. It is the library used to create and control the most of the graphical user interface: the menus, the toolbars, dialogs, child windows.

However, there is a very important part of the GUI which is not rendered using any library but by direct manipulation of the underlying HTML file: the sliding ribbon where the genome representation takes place. This kind of hand made manipulation is fundamental in order to get the most from the browser and it's very important to offer good response times and interactivity to the user.

The client has been build as a modular system with three modules with very clear functionality: the main GUI, the data proxy and the track-container modules. The modular design has important advantages when extending or adapting the system and allows for easy implementation of new functionalities.

Main GUI Module

The first module is the one in charge of the main window and GUI. It renders the main and track-container specific toolbars as well as dialogs and messages. It also controls the size of the screen and notifies the non-EXT parts of the application when a resizing is necessary via the Prototype custom-event system. The main GUI is what can be considered as the main module of the application since it creates and controls the others and is the main part interacting with the user. Most of the active user interface is build with EXT and its very convenient code based GUI creation functions. This module is also the one controlling the creation and destruction of track-containers and its positioning and sizing.

Data-Proxy Module

The second module is a data-proxy. It receives the data requests from the other client modules and request it from the server in a suitable way. It encapsulates and handles all AJAX communication between client and server and can be replaced or extended if other data sources or protocols should be used. This module is responsible of another fundamental feature of a system intended to manage and visualise large amounts of data: caching and data sharing.

The data-proxy stores all genomic data in the client and serves it to the different track-containers when required. The existence of a unique data-proxy ensures that the data used by different tracks in different track-containers is never requested more than once to the server reducing the server load, the client memory footprint and the network bandwith used.

Petitions to the data-proxy are made via its external interface but direct access to the underlying data structures is granted to the track-container modules in order to reduce the overhead of function calls in heavy used data. The notification of completed requests is made via the custom events system provided by Prototype.

Track-Container Module

The track-container is the module creating and handling all graphical representation of the genomic features and sequence. It handles the data requests to the data-proxy module as well as the user interaction with the data such as movements, zooms and addition or deletions of tracks.

Each track-container in the application is actually a long page loaded inside an HTML iframe. Each of these pages has its own copy of the track-container module running and controlling the specific genomic view. They are all linked via event observation of the main GUI module or the shared data-proxy module.

Two different kinds of tracks can be drawn in a track container:

Feature based tracks. This kind of track is rendered by creating a single and unique DIV HTML element for each feature in the track, and then attaching the necessary listeners to it. This respresentation allows for very fast response when dragging the genome and specially when zooming, since no new data have to be downloaded. The feature based tracks allows in client control of the representaction and easy customisation such feature colour changing, etc. Since in feature based tracks each feature has to be added individually to the DOM and the browser has to render them one by one, the exact representation of each feature is heavily optimised. Despite the optimisations, this representation is not suitable for very dense tracks with more than about ten or fifteen thousand features per entry point. This limit also varies depending on the client hardware and the browser used.

Image tracks. This second kind of track is used in dense tracks such as SNPs and basically any track with more than ten thousand features per entry point. In this case, the track is represented as a set of contiguous images representing each of them a small part of the genome at a given zoom level. In this case, the heavy work of finding the features present in a genomic region and its graphical representation is done by the server. Usually, the server not only creates the image itself but also an associated HTML map which is send to the browser alongside with a link to the image file in the JSON response to the AJAX request. This map is used to give to the images almost the same level of interactivity as the HTML based features – tooltip, extended information, useful links... Both the image and the associated map is cached

in the server in order to reduce the server load and the response time. Most current browsers are also configured to save a cached copy of images used in web pages and thus, when revisiting a region previously examined, most of the data will be already cached in the client, reducing latency and incrementing responsiveness. This kind of tracks is used to represent the underlaying sequence, which can be viewed at most close zoom levels.

The structure of the track-container as a long sliding ribbon is one of the keys to achieve speed and response. Simple small movements translate into movements of the whole ribbon reducing the repositioning computations to be made by the browser. When zooming, however, the tracks are completely redrawn from scratch for feature based tracks and new images are requested from the server for image based tracks.

The client has been developed and thoroughly tested using Firefox2, however it also been tested under Opera and Safari and it works correctly. The only browser where the client has problems is Internet Explorer and it's because of some quirks in its JavaScript handling.

4.2 The Server

The DASGenExp server subsystem is a collection of Perl scripts accessed via CGI interface. The main purpose of those scripts if fetching the data from the DAS sources and serving it to the client. However they also handle the application configuration and perform a very important task of caching.

Most of the interaction between the client and the server is done using the Bio::Das::Lite[7] Perl module. This module encapsulates the DAS queries in Perl functions getting and returning Perl objects. By using it all the hassle of parsing the XML – and eventually adapting to updates on the diverse DAS DTDs – is avoided.

When a client request the data for a track, the server connects to the DAS source and asks for all features of the requested type on the requested entry point. When the response is received, it is processed, transformed to JSON and send to the client. A copy of that data is also stored in the server cache and will be returned the next time the track is requested for that entry point. That scheme is used only in non-dense tracks – approximately les than ten thousand features per entry point – and thus responses can be fast.

In order to respond to an image request the server uses the cached data stored in a series of files corresponding to 10 kbases for sequence data and 10 Mbases for features. The images are created using the GD library and it's Perl bindings. The use of small files gives usually fast image creation and response despite in some cases the actual data have to be previously fetched from the actual DAS source. When generating images at a close enough zoom level – less than 200Kbp per 1000px – an associated HTML map is generated to associate feature information with the image. Both the images and maps are cached to avoid unnecessary server work.

The application configuration is made on the server by editing some simple text files and no access to any database is needed. Since it only requires an HTTP

server, a Perl interpreter and some freely available Perl packages, the custom install of DASGenExp is simple and straightforward. This option, however, is still not available since the application is still in development state.

5 Conclusions

We have developed DASGenExp, a web-based genome oriented DAS client easy to use and highly interactive. Our client gives to the user the opportunity to explore integrated data coming from different DAS sources using an intuitive GUI. It additionally offers some newer functionalities as showing multiple genomic locations at the same time, or arbitrary number of additional synchronized zoom windows. It is important to note that despite all the information it offers the user interface is very responsive specially when moving around the genome.

DASGenExp is currently available at `http://gralggen.lsi.upc.edu/recerca/DASgenexp/` and can be freely accessed.

6 Future Work

At the moment the server is running in a single Sun machine (Sun Fire 280R Server, 2xUltraSPARC III 1.2GHz, 4GB of RAM) and this is the cause of most of the lag that can be observed when downloading data and specially image based data. We plan to add more computational resources and we expect to get a big improvement in speed.

In the client part we are working on the possibility of the user dynamically adding new DAS sources to the sources list since the current list it is not exhaustive and there can even exist user specific sources.

Acknowledgements

This project is supported by research project LogicTools TIN2004-03382 and by a predoctoral FI grant from the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa of the Generalitat de Catalunya and the Euopean Social Fund.

References

1. Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R., Stein, L.: The Distributed Annotation System. BMC Bioinformatics 2, 7 (2001)
2. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The Human Genome Browser at UCSC. Genome Research 12(6), 996–1006 (2002)
3. Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H.R., Cox, A.V.: The Ensembl Web Site: Mechanics of a Genome Browser. Genome Research 14(5), 951–955 (2004)

4. Prlic, A., Down, T.A., Kulesha, E., Finn, R.D., Kahari, A., Hubbard, T.J.P.: Integrating sequence and structural biology with DAS. BMC Bioinformatics 8, 333 (2007)
5. Prototype JavaScript framework: Easy Ajax and DOM manipulation for dynamic web applications (May 2008), http://www.prototypejs.org/
6. Ext - A foundation you can build (May 2008), http://extjs.com/
7. Bio::Das::Lite - Perl extension for the DAS (HTTP+XML) Protocol (May 2008), http://biodas.org/, http://search.cpan.org/~rpettett/Bio-Das-Lite-1.054/lib/Bio/Das/Lite.pm

Data Integration Issues in the Reconstruction of the Genome-Scale Metabolic Model of *Zymomonas Mobillis*

José P. Pinto[1], Oscar Dias[2], Anália Lourenço[2], Sónia Carneiro[2], Eugénio C. Ferreira[2], Isabel Rocha[2], and Miguel Rocha[1]

[1] Departament of Informatics / CCTC
[2] IBB - Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga - Portugal
{josepedr,mrocha}@di.uminho.pt,
{odias,analia,soniacarneiro,ecferreira,irocha}@deb.uminho.pt

Abstract. Genome-scale model reconstruction represents a major tool in the field of Metabolic Engineering .This paper reports on a study about data integration issues in the process of genome-scale reconstruction of the metabolic model of the bacterium *Zymomonas mobilis,* a promising organism for bioethanol production. Data is retrieved from the Entrez Gene, KEGG, BioCyc and Brenda databases, and the several processes involved in data integration from these sources are described, as well as the data quality issues.

Keywords: Genome-scale model reconstruction, *Zymomonas mobillis*, data integration, data quality.

1 Introduction

Genome-scale reconstructed metabolic models are based on the well-known stoichiometry of biochemical reactions and can be used for simulating *in silico* the phenotypic behaviour of a microorganism under different environmental and genetic conditions, thus representing an important tool in metabolic engineering [1]. However, while the reconstruction of the metabolic network of an organism is likely to become a widespread procedure, starting with the fully sequenced and (partially) annotated genome sequence, it is currently far from being a standardized methodology [2]. This is due in part to the lack of uniform computational tools for model reconstruction, but primarily to the difficulties associated with the extraction of information other than what is available from the annotated genome.

In this paper, we address the reconstruction of the metabolic model of *Zymomonas mobilis* ZM4, among the most promising microorganisms for ethanol fuel production [3]. The genome-scale metabolic reconstruction is imperative for the feasibility of ongoing studies since there is no available genome-scale metabolic model for this organism. The number of reports in current literature studying its *in vivo* physiology remains small and there is a limited use of the metabolic engineering experimental and computational tools in the understanding of its metabolic pathway interconnectivity [4]. Therefore, genome-scale metabolic modeling stands out as one of the most promising approaches to obtain *in silico* predictions of cellular function based on the interaction of the cellular components [5,2].

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 92–101, 2009.
springerlink.com

This work is focused on the first steps of metabolic network reconstruction aiming at delivering valuable forms of automation that can assist on the collection and processing of the information required. This case study is invaluable, because of its importance as an ethanologenic source and the scarce availability of data to support related research. However, the workflow was planned in order to account for data issues in an organism-independent way. All the processes and analysis guidelines proposed can be applied to the reconstruction of models of other organisms, adjusting only data retrieval processes from particular repositories.

The main focus of this work lays on data integration planning, in particular on the assessment of data quality. Handling the diversity and quality of the contents along with data formats and structure is determinant to obtain a consistent repository. Most of the times, biologists rely on particular data sources, which they are familiar with. Understanding the reasoning that drives the expert while manually searching for data, allows the identification of the basic set of elements from each source and, far more important, how data sources can be linked together. From then on, source information extraction and preliminary processing can be fully automated and multi-source data integration can be achieved.

The establishment of a fully automated dataflow is inconceivable because data quality poses challenging issues that require expert non-trivial evaluation. Intra-source data quality is often disputable. Misspellings, nulls, duplicates and inconsistencies may undermine data acquisition and further integration. Multi-source integration raises additional quality concerns due to the scarce use of standard nomenclatures that raises terminological issues, namely term novelty, homonymy and synonymy. However, data processing can account for the most common issues, delivering descriptive quality-related statistics and proposing, when possible, candidate solutions to the integration and data quality issues.

2 Information Requirements

The genome-scale reconstruction of a metabolic network encompasses several steps [1], as depicted on Fig. 1. : (1) genome annotation; (2) identification of the biochemical reactions from the annotated genome sequence and available literature; (3) determination of the reaction stoichiometry including cofactor requirements; (4) definition of compartmentation and assignment of reaction localizations; (5) determination of the biomass composition; (6) measurement, calculation, or fitting of energy requirements; and (7) definition of additional constraints.

The process is laborious and requires substantial manual evaluation of the stoichiometry of different reactions in the network: whereas it typically takes 10% of the reconstruction time to collect 90% of all reactions from the annotated genome sequence, the remaining 90% of the time is spent collecting the remaining 10% of data from literature. The present work discusses the shadowed steps of the figure, namely the identification of reactions and collection of stoichiometric data.

For the microorganisms with fully sequenced genomes, the process of reconstructing the metabolic network starts with a careful inspection of the data obtained from the genome annotation. The process can be initiated by consulting a public repository

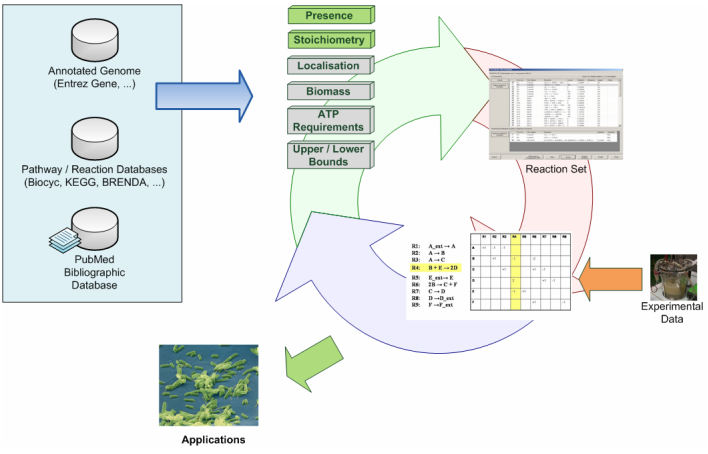


Fig. 1. An illustration of the metabolic network reconstruction process

of genome sequence data, such as GOLD [6], TIGR [7] or NCBI [8]. Important data to be extracted from these sources include gene or open reading frame (ORF) names, assigned cellular functions, sequence similarities, and, for enzyme encoding genes, the Enzyme Commission (EC) number(s) corresponding with the gene products. From the complete set of sequenced genes, only the genes encoding enzymes and membrane transporters are used for the reconstruction.

At the end of this process, the names of the genes assigned during genome annotation, the names of the reactions, reactants and products for each reaction should all be included in the reaction list. Typically, the initial reconstruction only considers genes that code for enzymes with EC numbers assigned. Public pathway databases, such as BRENDA [9] or KEGG [10] provide detailed information about each individual reaction catalyzed by a enzyme with assigned EC number.

Then, the reaction set has to be complemented with reactions catalyzed by enzymes that do not have EC numbers assigned, with transport and exchange reactions, and with reactions known to exist in a given organism, but for which no corresponding genes have been found during annotation. This can only be accomplished by thorough curation of publications and biochemistry textbooks. Curation may be fully manual or comprise the use of Biomedical Text Mining techniques. Either way, due to terminological issues and the challenges posed by unstructured text processing, it is laborious and time-consuming.

Despite the many obstacles faced, this information validates the data deduced from the genome and discarding questionable reactions with poor annotation based on low sequence similarity and those for which no evidence has been found in literature. Also, it supports the selection of reaction(s) specific to the organism being reconstructed from the multiple potential reactions associated with each given EC number in public databases. Furthermore, special cases with more complex than one-gene-to-one-enzyme-to-one-reaction relations need to be considered: (1) many enzymes accept several different substrates; (2) isoenzymes are encoded by different genes, but each of them catalyzes the same reaction(s); (3) for reactions catalyzed by enzyme complexes, several genes are associated with one or more reactions [1]. Information

about reaction stoichiometry can also be found in public databases only for enzymes with assigned EC numbers. For all other reactions, stoichiometric information should be based on the literature data.

3 Data Integration

Our workflow is illustrated in Figure 2 and encompasses the following steps: (1) data loading from original sources into temporary tables; (2) single-source debugging; (3) single-source quality-related processing; (4) detection of conflicts on multi-source integration; (5) semi-automatic conflict resolution; (6) multi-source contents integration; and (7) enforcement of data quality.

3.1 Data Source Description

The parameters related to genome annotation were taken from the NCBI's Entrez Gene [11]. The list of reactions and stoichiometry data was delivered by integrated contents from KEGG, BioCyc and Brenda.

The Kyoto Encyclopedia of Genes and Genomes (**KEGG**) is an information repository that contains several kinds of biological data, with the purpose of linking genomic information with higher order functional information [12]. KEGG is composed by 19 highly integrated databases, each one belonging to one of three categories: Systems, Genomic or Chemical [13]. Due to this widespread hyperlinking, KEGG was considered one of the "central" data sources of this project. KEGG's data is organized in organism-specific and class-specific subdirectories where data is kept in text files. Currently, we extract information from databases of the genomic and chemical categories and also information about pathways.

BRaunschweig ENzyme DAtabase (**BRENDA**) - is the main collection of enzyme functional data available and its data is primarily collected from literature [14]. We use it to complement the enzyme data extracted from generic data sources (for example KEGG or BioCyc). BRENDA is a very extensive database characterized by the fact that it is not limited to specific organisms or aspects of enzymes, covering a wide range of information for all enzymes [15].

In BRENDA, the information is organized by EC number, and within each EC number, it is further organized by organism and by the documents from which it was extracted. One organism can have more than one enzyme with the same EC number, but since in the primary literature EC numbers are rarely associated to specific sequences, discriminating between the enzymes with the same EC number is not possible [16]. BRENDA's contents are delivered in two text files.

BioCyc is a collection of Pathway/Genome Databases (PGDBs), quite popular among biological researchers [17]. It is a very extensive data source containing more than 160 different PGDBs, each one covering a specific organism. BioCyc repositories contain information about most eukaryotic and prokaryotic species whose genome has been sequenced [18]. BioCyc is available in the form of database dumps or as flat files (the latter was chosen).

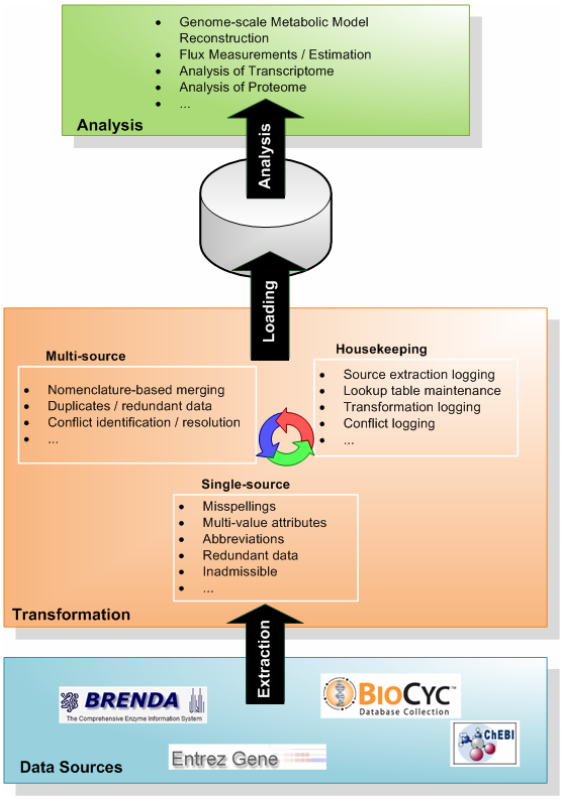


Fig. 2. Workflow for the data integration process

3.2 Data Quality Issues

Data quality issues can be classified into single-source and multi-source issues and, within these, schema and instance-related issues. The quality of a data source depends on schema and integrity constraints. Sources without a schema (e.g. flat files) raise a higher number of errors and inconsistencies. Database systems are expected to enforce a data model and application-specific integrity constraints. Usually, schema-related problems (e.g. uniqueness and referential violations) occur due to the lack of integrity constraints, data model limitations and poor schema design or because integrity constraints were limited to prevent control overhead.

Single-source quality issues get worse when performing data integration. Each source may contain dirty data and the data in the sources may be represented differently, overlap or contradict. At the schema level, data model and schema design differences are to be addressed by the steps of schema translation and schema integration, respectively. Naming conflicts arise when the same name is used for different entities (homonyms) or different names are used for the same entity (synonyms). Structural conflicts occur in many variations and refer to different representations of the same entity, different component structure, different data types, different integrity

constraints, etc. In addition to schema-level conflicts, data problems from single sources can occur with different representations in different sources (e.g., duplicated or contradicting records).

KEGG. The initial analysis of KEGG (Table 1) revealed some issues in the compound data, namely some data such as mass and formula are missing. However, the percentage of records with missing data is relatively small.

Table 1. Characteristics of the KEGG compound data for *Zymomonas mobilis*

Number of compounds	15050
Compounds without formula	2322
Compounds without mass	3457
Compounds without formula & mass	2322

In KEGG, it is considered that in a organism there is only one enzyme for each EC number. This fact makes it difficult to integrate KEGG with data sources that have more detailed enzymatic information (like BioCyc). Also, the study of reaction-enzyme associations (Table 2) revealed that some enzymes in catalogue are not associated with any reaction. One serious problem in KEGG is that pathway data is stored in an image format, hard to parse, and consequently, to combine with other formats of data. This situation appears to be changing however since the KEGG pathway data is being migrated into an XML format (KGML).

Table 2. Characteristics of the KEGG enzymatic data for *Zymomonas mobilis*

Number of enzymes	413
Number of reactions	942
Number of enzymes associated with reactions	380

BRENDA. Two main difficulties were found during BRENDA's data processing: (1) EC numbers are formal record identifiers, but often the identifier field has comments that constrain the integration of the records; (2) the identification of the compounds that affect the enzymes (e.g. inhibitors, cofactors and activators) is only provided by name. After analyzing EC number field values, it was possible to establish a parsing schema that allows adequate record crossing. Compound name resolution is far more delicate and brings a high error probability to the integration. Since the names are the only identifiers available there was no other choice but to use them as the basis for the integration. In order to verify the viability of the integration, a terminology comparison study was undertaken (Tables 3 and 4).

This study showed that there are relatively few name conflicts when the BRENDA compound data is compared with KEGG and BioCyc. Consequently, the integration of the BRENDA data with the information from these two databases should not be a problem. However, KEGG and BioCyc are not databases specific for compounds, so

Table 3. Number of compounds in BRENDA entries about *Zymomonas mobilis* and the number of compounds successfully associated/number of conflicts found during the association of the BRENDA compounds with information from other data sources

Compounds found / Conflicts	N compounds	KEGG	BioCyc	CHEBI
BRENDA cofactors	13	13 / 0	7 / 0	8 / 5
BRENDA inhibitors	52	37 / 3	15 / 0	13 / 24
BRENDA activating compounds	11	5 / 0	1 / 0	2 / 4

there is no guarantee that all or even most of the possible names for a compound are present. For this reason, the BRENDA data was also compared with ChEBI, a compound specific database [19]. The larger number of conflicts obtained with ChEBI leads to the conclusion that the integration of the compound data from BRENDA with the one extracted from other data sources, given the redundancy of the names of the compounds, is far more difficult that it was originally expected. In fact, the only way of insuring a correct integration is manual curation. Another potentially serious problem with data obtained from BRENDA is the fact that in this database there is only a small quantity of information about *Zymomonas mobilis*, specifically only 29 entries.

BIOCYC. The greatest obstacles to the integration of the BioCyc and KEGG genomic data is the scarce use of standard identifiers. The only common gene is the name of the genes and this identifier, similarly to the name of the compounds, is subject to a great degree of redundancy. Furthermore, only some of the genes have names associated in the database. In fact only 663 out of 1998 genes can be integrated (Table 4). Since there is no way to solve the problem only with the information from BioCyc and KEGG, the possibility of using a third database should be considered, preferably with links to both BioCyc and KEGG.

Table 4. Characteristics of the BioCyc genomic data for *Zymomonas mobilis*

Total number of Genes	1998
Genes with no redundant name	663
Genes with redundant names	54

Another problem with the BioCyc data is the fact that the EC number is not associated with the enzymes but rather with reactions. This problem is solved, since when a reaction has an EC number it can be considered that all enzymes that catalyze that reaction have that EC number. This method allows the association of an EC number to most enzymes identified in the BioCyc data (Table 5).

The integration of the KEGG and BioCyc pathway data presents another challenge, because the information is kept in quite different formats. In KEGG, this data is kept as images with hyperlinks in certain regions and in BioCyc the data is stored in text format. Because of their differences, it is not possible to integrate both types of data and it will be necessary to include both or choose only one. There was one more difficulty with the BioCyc database: the references to external database sources in BioCyc

Table 5. Characteristics of protein and enzyme data from BioCyc for *Zymomonas mobilis*

Proteins	2007
Enzymes	610
Reactions	880
Enzyme-reaction associations	837
Reactions with enzymes	636
Enzymes to witch may be associated an encumber	542

are stored in an unusual format: (PID "56544468" NIL |kaipa| 3390578134 NIL NIL). It was found that the first word is a code for the database and the second is the database id. Since there is not data that associates the database code to the corresponding database in BioCyc the association has to be done manually (through consulting the BioCyc site) before integrating the data. Fortunately, there are only a few databases codes in BioCyc.

3.3 Integration Strategy

The data integration strategy is divided in three stages, each corresponding to the integration of one of the data sources into a shared database, the Data Staging Area (DSA). The data sources are loaded to the DSA in a specific order: KEGG first, then BioCyc and finally BRENDA. When information from a data source is added into the DSA the data is compared with the ones already present in the DSA. It is then added if it is not present or used to complement the information already in the DSA in the case that part of the new information is already present. The new data is compared with the data already in the DSA normally by the use of non redundant identifiers (e.g CAS number for compounds) or by the comparison of a set of factors that combined can be used as a nearly non redundant identifier.

The loading of the KEGG information into the DSA is the first step. There is a problem in the KEGG data that affects the integration of BioCyc: in KEGG it is considered that only one enzyme is associated to one EC number. This problem was solved by observing that in KEGG genes, the product of the gene is identified, including its EC number (in the case it is an enzyme). To overcome this difficulty, we cross the information from the KEGG enzyme and gene data, to identify the individual enzymes associated with an EC number. After solving this issue, the remaining of the loading is handled smoothly.

The integration of BioCyc starts with the compound data, since it is the easiest to integrate, followed by the gene and the protein data. This is a two step process: in the first phase, all the proteins are compared with the ones already in the DSA, through the genes that code them to determine which ones are added and which are complemented; the second step is the identification of the new enzymes that can only be executed during the integration of the reaction data. Next, the reaction data is integrated. The reactions are not directly compared with the ones already in the DSA, instead they are compared with the enzymes that catalyze them and their reactants and products. The process ends with the integration of the pathway data.

The integration of BRENDA starts by comparing the EC numbers in its entries with the ones in the DSA; when the values match, the BRENDA data is associated

with that record. BRENDA data includes references to compounds that affect the reaction, such as inhibitors or cofactors. Determining if these compounds are already in the DSA or if they must be added is difficult, since the only identification is the name of the compound. Since there can be multiple enzymes associated with EC numbers, this means that the same BRENDA data will probably be linked with the same enzymes.

4 Conclusions

In this work, we approached a number of issues related to the process of genome-scale reconstruction of the metabolic model of the bacterium *Zymomonas mobilis.* These were mainly related to data quality issues and the implementation of suitable data integration processes. A number of problems were identified and useful guidelines for their solution were proposed. This work is on-going and it will proceed by enlarging the set of handled conflicts and integrating other data sources.

Acknowledgements

The authors acknowledge the support from the Portuguese FCT under the project POCI/BIO/60139/2004 and the PhD grant (ref. SFRH/BD/41763/2007).

References

1. Rocha, I., Forster, J., Nielsen, J.: Design and application of genome-scale reconstructed metabolic models. Methods Mol. Biol. 416, 409–431 (2008)
2. Notebaart, R.A., van Enckevort, F.H., Francke, C., Siezen, R.J., Teusink, B.: Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics (7), 296 (2006)
3. Seo, J.S., Chong, H., Park, H.S., Yoon, K.O., Jung, C., Kim, J.J., Hong, J.H., Kim, H., Kim, J.H., Kil, J.I., Park, C.J., Oh, H.M., Lee, J.S., Jin, S.J., Um, H.W., Lee, H.J., Oh, S.J., Kim, J.Y., Kang, H.L., Lee, S.Y., Lee, K.J., Kang, H.S.: The genome sequence of the ethanologenic bacterium Zymomonas mobilis ZM4. Nat. Biotechnol. 1(23), 63–68 (2005)
4. Tsantili, I.C., Karim, M.N., Klapa, M.I.: Quantifying the metabolic capabilities of engineered Zymomonas mobilis using linear programming analysis. Microb.Cell Fact. (6), 8 (2007)
5. Borodina, I., Nielsen, J.: From genomes to in silico cells via metabolic networks. Curr. Opin. Biotechnol. 3(16), 350–355 (2005)
6. GOLD (Genomes OnLine Database v 2.0) web site (2008), `http://www.genomesonline.org/`
7. TIGR web site. TIGR web site (2008)
8. NCBI web site. NCBI web site (2008)
9. BRENDA web site. BRENDA web site (2008)
10. KEGG (Kyoto Encyclopedia of Genes and Genomes) web site (2008), `http://www.genome.jp/kegg/`
11. NCBI's Entrez Gene Web site (2008), `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`

12. Minorou, K., Susumu, G.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 1(28), 27–30 (2000)
13. Minorou, K., Michihiro, A., Susumu, G., Masahiro, H., Mika, H., Masumi, I., Toshiaki, K., Shuichi, K., Shujito, O., Toshiaki, T., Yoshihiro, Y.: KEGG for linking genomes to life and the environment. Nucleic Acids Research 36 (2007)
14. BRENDA introduction. BRENDA web site (2008)
15. Schomburg, I., Chanf, A., Hofmann, O., Ebeling, C., Ehrentreich, F., Schomburg, D.: BRENDA: a resource for enzyme data and metabolic information. TRENDS in Biochemical Sciences 1(27), 54–56 (2002)
16. Schomburg, I., Chang, A., Schomburg, D.: BRENDA, enzyme data and metabolic information. Nucleic Acids Research 1(30), 47–49 (2001)
17. BioCyc Introduction. BioCyc web page (2008)
18. Karp, P., Ouzounis, C., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., López-Bigas, N.: Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research 19(33) (2005)
19. CHEBI web site. CHEBI web site (2008)

Applying CBR Systems to Micro Array Data Classification

Sara Rodríguez, Juan F. De Paz, Javier Bajo, and Juan M. Corchado

Departamento de Informática y Automática, Universidad de Salamanca Plaza de la Merced s/n, 37008, Salamanca, Spain
{srg,fcofds,jbajope,corchado}@usal.es

Summary. Microarray technology allows to measureing the expression levels of thousands of genes in an experiment. This technology required requires computational solutions capable of dealing with great amounts of data and as well as techniques to explore the data and extract knowledge which allow patients classification. This paper presents a systems based on Case-based reasoning (CBR) for automatic classification of leukemia patients from microarray data. The system incorporates novel algorithms for data mining that allow to filter and classify as well as extraction of knowledge. The system has been tested and the results obtained are presented in this paper.

Keywords: Case-based Reasoning, HG U133, dendogram, leukemia classification, decision tree.

1 Introduction

The progress in the biomedicine [1] [24] and the incorporation of computational and artificial intelligence techniques, have caused a big progress in the detection and diagnosis of many illness. Microarray technology allows to measure the expression levels of thousands of genes in an experiment. This technology has been adopted by the research community for the study of a wide range of biologic processes allowing carry out diagnosis. Currently, it is being very used [23] for diagnosing of cancer such as Leukemias. This technique studies RNA chains thereby identifying the level of expression for each gene studied. It consists of hybridizing a sample for a patient and colouring the cellular material with a special dye. This offers different levels of luminescence that can be analyzed and represented as a data array [24]. These methods and tools need to work with expression arrays containing a large amount of data points. Specifically, the HG U133 plus 2.0 are chips used for this kind of analysis. These chips analyze the expression level of over 47.000 transcripts and variants, including 38.500 well-characterized human genes. It is comprised of more than 54.000 probe sets and 1.300.000 distinct oligonucleotide Feature. The HG U133 plus 2.0 provides multiple, independent measurements for each transcript. Multiple probes mean you get a complete data set with accurate, reliable, reproducible results from every experiment. Eleven pairs of oligonucleotide probes are used to measure the level

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 102–111, 2009.
springerlink.com

of transcription of each sequence represented on the GeneChip Human Genome Focus Array.

The process of studying a microarray is called expression analysis and consists of a series of phases: data collection, data preprocessing, statistical analysis, and biological interpretation. These phases analysis consists basically of three stages: normalization and filtering; clustering and classification; and extraction of knowledge. These stages can be automated and included in a CBR [15] system. The first step is critical to achieve both a good normalization of data and an initial filtering to reduce the dimensionality of the data set with which to work [5]. Moreover, the choice of a clustering technique allows data to be grouped according to certain variables that dominate the behaviour of the group [6]. After organizing into groups it is possible to extract the information necessary about the most significant probes that characterize each cluster. In base on this information, the association of new individuals to a cluster can be carried out. Finally, experts can learn from the analysis process.

For some time now, we have been working on the identification of techniques to automate the reasoning cycle of several CBR systems applied to complex domains [22] [15]. The microarray analysis to distinguish subclasses in disease and identify pattern associated with disease according to its genes. This patterns of expression that are used to classify types leukimia. This paper presents a CBR system that facilitates the analysis and classification of data from microarrays corresponding to patients with leukemia. Leukemia, or blood cancer, is a disease that has a significant potential for cure if detected early [4]. The model aims to improve the cancer classification based on microarray data using CBR. The system presented in this paper uses a model which takes advantage of three methods for analyzing microarray data: a technique for filtering data, a technique for clustering and a method for extracting the knowledge.

The paper is structured as follows: The next section presents the problem that motivates this research, i.e., the classification of leukemia patients from samples obtained through microarrays. Section 2 and Section 3 describe the proposed CBR model and how it is adapted to the problem under consideration. Finally, Section 4 presents the results and conclusions obtained after testing the model.

2 CBR System for Classifying Micro Array Data

The CBR developed tool receives data from the analysis of chips and is responsible for classifying of individuals based on evidence and existing data. Case-based Reasoning is a type of reasoning based on the use of past experiences [7]. CBR systems solve new problems by adapting solutions that have been used to solve similar problems in the past, and learning from each new experience. The primary concept when working with CBRs is the concept of case. A case can be defined as a past experience, and is composed of three elements: A problem description, which delineates the initial problem; a solution, which provides the sequence of actions carried out in order to solve the problem; and the final stage, which describes the state achieved once the solution was applied. A CBR

manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain. The retrieve phase starts when a new problem description is received. In this phase a similarity algorithm is used to find the greatest number of cases in the cases memory. In our case study, it conducted a filtering of variables, recovering important variables of the cases to determine the most influential in the conduct classification as explained in section 2.1. Once the most important variables have been retrieved, the reuse phase begins, adapting the solutions for the retrieved cases to obtain the clustering. Once this grouping is accomplished, the next step is to determine the provenance of the new individual to be evaluated. The revise phase consists of an expert revision for the solution proposed, and finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases, consequently updating the cases memory.

2.1 Retrieve

Contrary to what usually happens in the CBR, our case study is unique in that the number of variables is much greater than the number of cases. This leads to a change in the way the CBR functions so that instead of recovering cases at this stage, important variables are retrieved. Traditionally, only the similar cases to the current problem are recovered, often because of performance, and then adapted. In the case study, the number of cases is not the problem, rather the number of variables. For this reason variables are retrieved at this stage and then, depending on the identified variables, the other stages of the CBR are carried out. This phase will be broken down into 5 stages which are described below:

RMA: The RMA (Robust Multi-array Average) [8] algorithm is frequently used for pre-processing Affymetrix microarray data. RMA consists of three steps: (i) Background Correction; (ii) Quantile Normalization (the goal of which is to make the distribution of probe intensities the same for arrays); and (iii) Expression Calculation: performed separately for each probe set n. To obtain an expression measure we assume that for each probe set n, the background adjusted, normalized and log transformed intensities, denoted with Y, follow a linear additive model

$$x_{i,j,n} = \mu_{i,n} + \alpha_{j,n} + \epsilon_{i,j,n} \; with \; i = 1...I, j = 1...J, n = 1...N \sum_j \alpha_j = 0 \quad (1)$$

Where α_j is a probe affinity effect, μ_j represents the log_2 scale expression level for array i and $\epsilon_{i,j}$ represents an independent identically distributed error term with mean 0. Median polish [21] is used to obtain estimates of the values.

Control and error: During this phase, all probes used for testing hybridization are eliminated. These probes have no relevance at the time when individuals are classified, as there are no more than a few control points which should contain the same values for all individuals. If they have different values, the case should

be discarded. Therefore, the probes control will not be useful in grouping individuals. On occasion, some of the measures made during hybridization may be erroneous; not so with the control variables. In this case, the erroneous probes that were marked during the implementation of the RMA must be eliminated.

Variability: Once both the control and the erroneous probes have been eliminated, the filtering begins. The first stage is to remove the probes that have low variability. This work is carried out according to the following steps:

1. Calculate the standard deviation for each of the probes j

$$\sigma_{.j} = +\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\bar{\mu_{.j}} - x_{ij}\right)^2} \tag{2}$$

   Where N is the number of items total, is the average population for the variable j, is the value of the probe j for the individual i.
2. Standardize the above values

$$z_i = \frac{\sigma_{.j} - \mu}{\sigma} \tag{3}$$

   where $\mu = \frac{1}{N}\sum_{j=1}^{N}\sigma_{.j}$ and $\sigma_{.j} = +\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\bar{\mu_{.j}} - x_{ij}\right)^2}$ where $z_i \equiv N(0,1)$
3. Discard of probes for which the value of z meet the following condition: $z < -1.0$ given that $P(z < -1.0) = 0.1587$. This will effect the removal of about 16% of the probes if the variable follows a normal distribution.

Uniform Distribution: Finally, all remaining variables that follow a uniform distribution are eliminated. The variables that follow a uniform distribution will not allow the separation of individuals. Therefore, the variables that do not follow this distribution will be really useful variables in the classification of the cases. The contrast of assumptions followed is explained below, using the Kolmogorov-Smirnov [13] test as an example.

$$D = max\left\{D^+, D^-\right\} \tag{4}$$

where $D^+ = max_{1\leq i\leq n}\left\{\frac{i}{n} - F_o\left(x_i\right)\right\}$ $D^- = max_{1\leq i\leq n}\left\{F_o\left(x_i\right) - \frac{i-1}{n}\right\}$ with i as the pattern of entry, n the number of items and $F_o\left(x_i\right)$ the probability of observing values less than i with H_o being true. The value of statistical contrast is compared to the next value:

$$D_\alpha = \frac{C_\alpha}{k\left(n\right)} \tag{5}$$

in the special case of uniform distribution $k\left(n\right) = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$ and a level of significance $\alpha = 0.05$ $C_\alpha = 1.358$.

Correlations: At the last stage of the filtering process, correlated variables are eliminated so that only the independent variables remain. To this end, the linear correlation index of Pearson is calculated and the probes meeting the following condition are eliminated.

$$r_{x_{\cdot i}y_{\cdot j}} > \alpha \tag{6}$$

being: $\alpha = 0.95$ $r_{x_{\cdot i}y_{\cdot j}} = \dfrac{\sigma_{x_{\cdot i}y_{\cdot j}}}{\sigma_{x_{\cdot i}}\sigma_{y_{\cdot j}}}$ $r_{x_{\cdot i}y_{\cdot j}} = \dfrac{1}{N}\sum_{s=1}^{N}(\bar{\mu}_{\cdot i} - x_{si})(\bar{\mu}_{\cdot j} - x_{sj})$ where $r_{x_{\cdot i}y_{\cdot j}}$ is the covariance between probes i and j.

2.2 Reuse

Once filtered and standardized the data using different techniques of data mining, the system produce a set of values x_{ij} with i = 1 ... N, j = 1 ... s where N is the total number of cases, s the number of end probes. The next step is to perform the clustering of individuals based on their proximity according to their probes. Since the problem on which this study is based contained no prior classification with which training could take place, a technique of unsupervised classification was used. There is a wide range of possibilities in data mining. Some of these techniques are artificial neural networks such as SOM [9] (self-organizing map), GNG [10] (Growing neural Gas) resulting from the union of techniques CHL [11] (Competitive Hebbian Learning) and NG [12] (neural gas), GCS [10] (Growing Cell Structure). There are other techniques with less computational cost that provide efficient results. Among them we can find the dendogram and the PAM method [16] (Partitioning Around Medoids). A dendrogram [17] is a ascendant hierarchical method with graphical representation that facilitates the interpretation of results and allows an easy way to establish groups without prior establishment. The PAM method requires a selection of the number of clusters previous to its execution.

The dendograms are hierarchical methods that initially define as conglomerates for each available cases. At each stage the method joins those conglomerates of smaller distance and calculates the distance of the conglomerate with everyone else. The new distances are updated in the matrix of distances. The process finishes when there is one only conglomerate (agglomerative method). The distance metric used in this paper has been the average linkage. This metric calculates the average distance of each pair of nodes for the two groups, and based on these distances mergers the groups. The metric is known as unweighted pair group method using arithmetic averages (UPGMA) [18]. Once the dendogram has been generated, the error rate is calculated bearing in mind the previous cases. If the accuracy rate is up to 80%, the extraction of knowledge using the CART (Classification and Regression Tree) [19] algorithm is carried out, and finally the new case is classified. The CART algorithm is a non parametric test that allows extracting rules that explain the classification carried out in the previous steps. There are others techniques to generate the decision trees, that is the case of the methods based on ID3 trees [20], although the most used

currently is CART. This method allows to generate rules and to extract the most important variables to classify patients with high performance.

2.3 Revise and Retain

The revision is carried out by an expert who determines the correction with the group assigned by the system. If the assignation is considered correct, then the retrieve and reuse phases are carried out again so that the system is ready for the next classification

3 Case Study

In the case study presented in the framework of this research are available 232 samples are available from analyses performed on patients either through punctures in marrow or blood samples. The aim of the tests performed is to determine whether the system is able to classify new patients based on the previous cases analyzed and stored.

Figure 1 shows a scheme of the bio-inspired model intended to resolve the problem described in Section 2. The proposed model follows the procedures that are performed in medical centres. As can be seen in Figure 1, a previous phase, external to the model, consists of a set of tests which allow us to obtain data from the chips and are carried out by the laboratory personnel. The chips are hybridized and explored by means of a scanner, obtaining information on the marking of several genes based on the luminescence. At that point, the CBR-based model starts to process the data obtained from the microarrays.

The retrieve phase receives an array with a patient's data as input information. It should be noted that there is no filtering of the patients, since it is the work

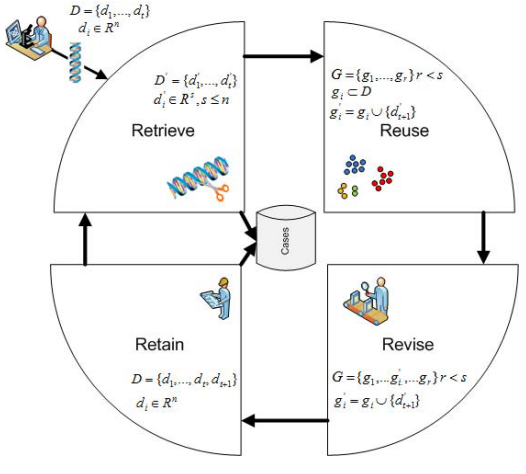


Fig. 1. Proposed CBR model

of the researcher conducting this task. The retrieve step filters genes but never patients. The aim of this phase is to reduce the search space to find data from the previous cases which are similar to the current problem. The set of patients is represented as $D = \{d_1, ..., d_t\}$, where $d_i \in \mathbb{R}^n$ represents the patient i and n represents the number of probes taken into consideration. As explained in Section 2.1 during the retrieve phase the data are normalized by the RMA algorithm [8] and the dimensionality is reduced bearing in mind, above all, the variability, distribution and correlation of probes. The result of this phase reduces any information not considered meaningful to perform the classification. The new set of patients is defined through s variables $D' = \left\{d'_1, ..., d'_t\right\}$ $d'_i \in \mathbb{R}^s, s \leq n$.

The reuse phase uses the information obtained in the previous step to classify the patient into a leukemia group. The patients are first grouped into clusters. The data coming from the retriever phase consists of a group of patients $D' = \left\{d'_1, ..., d'_t\right\}$ with $d'_i \in \mathbb{R}^s, s \leq n$ each one characterized by a set of meaningful attributes $d_i = (x_{i1}, ..., x : is)$, where x_{ij} is the luminescence value of the probe i for the patient j. In order to create clusters and consequently obtain patterns to classify the new patient, the reuse phase implements a method of hierarchical cluster called dendogram, which has been explained in section 2.2. The system classifies the patients by taking into account their proximity and their density, in such a way that the result provided is a set G where $G = \{g_1, ..., g_r\}$ $r < s$ $g_i \subset D$, $g_i \cap g_j = \phi$ with $i \neq j$ and $i, j < r$. The set G is composed of a group of clusters, each of them containing patients with a similar disease. The clusters have been constructed by taking into account the similarity between the patient's meaningful symptoms. Once the clusters have been obtained, the accuracy rate is calculated, if it is greater than 80% then the clustering and extraction of knowledge are carried out. The new patient is defined as d'_{t+1} and his membership to a group is determined following the classification tree in section 2.2. The result of the reuse phase is a group of clusters $G = \left\{g_1, ..., g'_i, ..., g_r\right\}$ $r < s$ where $g'_i = g_i \cup \left\{d'_{t+1}\right\}$.

An expert from the Cancer Institute is in charge of the revision process. This expert determines if $g'_i = g_i \cup \left\{d'_{t+1}\right\}$ can be considered as correct. In the retain phase the system learns from the new experience. If the classification is considered successful, then the patient is added to the memory case $D = \{d_1, ..., d_t, d_{t+1}\}$.

4 Results and Conclusions

This paper has presented a CBR system which allows automatic cancer diagnosis for patients using data from microarrays. The model combines techniques for the reduction of the dimensionality of the original data set and a method of clustering and extraction the knowledge. The system works in a way similar to how human specialists operate in the laboratory, but is able to work with great amounts of data and make decisions automatically, thus reducing significantly

both the time required to make a prediction, and the rate of human error due to confusion. The CBR system presented in this work focused on identifying the important variables for each of the variants of blood cancer so that patients can be classified according to these variables.

In the study of leukemia on the basis of data from microarrays, the process of filtering data acquires special importance. In the experiments reported in this paper, we worked with a database of bone marrow cases from 212 adult patients with five types of leukaemia. The retrieve stage of the proposed CBR system presents a novel technique to reduce the dimensionality of the data. The total number of variables selected in our experiments was reduced to 785, which increased the efficiency of the cluster probe. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Institute of Cancer. The error rates have remained fairly low especially for cases where the number of patients was high. To try to increase the reduction of the dimensionality of the data we applied principal components (PCA) [14], following the method of Eigen values over 1. A total of 93 factors were generated, collecting 96% of the variability. However, this reduction of the dimensionality was not appropriate in order to obtain a correct classification of the patients. Figure 2a shows the classification performed for patients from all the groups. In the left it is possible to observe the groups identified in the classification process. Cases interspersed represent individuals with different classification to the previous-one. As shown in Figure 2a the number of misclassified individuals have been low.

Once checked that the retrieved probes allow classifying the patients in similar way to the original one, we can conclude that the retrieve phase works satisfactorily. Then, the extraction of knowledge is carried out bearing in mind the selected probes. The algorithm used was CART [19] and the results obtained are shown in Figure 2b.

The proposed model resolves this problem by using a technique that detects the genes of importance for the classification of diseases by analysing the available data. As demonstrated, the proposed system allows the reduction of the

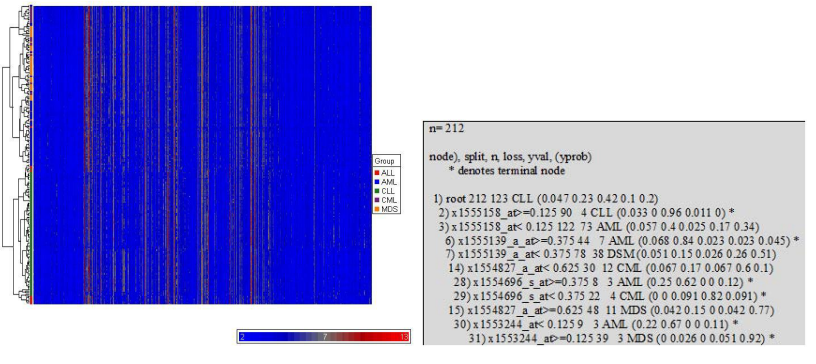


Fig. 2. Classification obtained

dimensionality based on the filtering of genes with little variability and those that do not allow a separation of individuals due to the distribution of data. It also presents a technique for clustering based in hierarchical methods. The results obtained from empirical studies are promising and highly appreciated by specialists from the laboratory, as they are provided with a tool that allows both the detection of genes and those variables that are most important for the detection of pathology, and the facilitation of a classification and reliable diagnosis, as shown by the results presented in this paper.

Acknowledgments

Special thanks to the Institute of Cancer for the information and technology provided.

References

1. Shortliffe, E., Cimino, J.: Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, Heidelberg (2006)
2. Tsoka, S., Ouzounis, C.: Recent developments and future directions in computational genomics. FEBS Letters 480(1), 42–48 (2000)
3. Lander, E., et al.: Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001)
4. Rubnitz, J., Hijiya, N., Zhou, Y., Hancock, M., Rivera, G., Pui, C.: Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia. Pediatric Blood & Cancer 44(2), 138–141 (2005)
5. Armstrong, N., van de Wiel, M.: Microarray data analysis: From hypotheses to conclusions using gene expression data. Cellular Oncology 26(5-6), 279–290 (2004)
6. Quackenbush, J.: Computational analysis of microarray data. Nature Review Genetics 2(6), 418–427 (2001)
7. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (1993)
8. Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T.: Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data. Biostatistics 4, 249–264 (2003)
9. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, 59–69 (1982)
10. Fritzke, B.: A growing neural gas network learns topologies. In: Tesauro, G., Touretzky, D., Leen, T. (eds.), Advances in Neural Information Processing Systems, vol. 7, pp. 625–632, Cambridge (1995)
11. Martinetz, T.: Competitive Hebbian learning rule forms perfectly topology preserving maps. In: ICANN 1993: International Conference on Artificial Neural Networks, pp. 427–434. Springer, Heidelberg (1993)
12. Martinetz, T., Schulten, K.: A neural-gas network learns topologies. In: Kohonen, T., Makisara, K., Simula, O., Kangas, J. (eds.) Artificial Neural Networks, Amsterdam, pp. 397–402 (1991)
13. Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition 34, 1625–1637 (2001)
14. Jolliffe, I.: Principal Component Analysis, 2nd edn. Series in Statistics. Springer, Heidelberg (2002)

15. Riverola, F., Daz, F., Corchado, J.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence 22(3-4), 254–268 (2006)
16. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
17. Saitou, N., Nie, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425 (1987)
18. Sneath, P., Sokal, R.: Numerical Taxonomy. The Principles and Practice of Numerical Classification. W.H. Freeman Company, San Francisco (1973)
19. Breiman, L., Friedman, J., Olshen, A., Stone, C.: Classification and regression trees. Wadsworth International Group, Belmont (1984)
20. Quinlan, J.: Discovering rules by induction from large collections of examples. In: Michie, D. (ed.) Expert systems in the micro electronic age, pp. 168–201. Edinburgh University Press, Edinburgh (1979)
21. Holder, D., Raubertas, R., Pikounis, V., Svetnik, V., Soper, K.: Statistical analysis of high density oligonucleotide arrays: a SAFER approach. In: Proceedings of the ASA Annual Meeting Atlanta, GA (2001)
22. Corchado, J., Corchado, E., Aiken, J., Fyfe, C., Fdez-Riverola, F., Glez-Bedia, M.: Maximum Likelihood Hebbian Learning Based Retrieval Method for CBR Systems. In: Proceedings. of the 5th International Conference on Case-Based Reasoning, pp. 107–121 (2003)
23. Quackenbush, J.: Microarray Analysis and Tumor Classification. The new england journal o f medicine, 2463–2472 (2006)
24. Zhenyu, C., Jianping, L., Liwei, W.: A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artificial Intelligence in Medicine 41, 161–175 (2007)

Multiple-Microarray Analysis and Internet Gathering Information with Application for Aiding Medical Diagnosis in Cancer Research

Daniel Glez-Peña[1], Manuel Glez-Bedia[2], Fernando Díaz[3], and Florentino Fdez-Riverola[1]

[1] Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
{dgpena,riverola}@uvigo.es
[2] Dept. de Informática e Ingeniería de Sistemas, University of Zaragoza, Edificio Ada Byron, María de Luna, 1, 50018, Zaragoza, Spain
mgbedia@unizar.es
[3] Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
fdiaz@infor.uva.es

Abstract. In light of the fast growth in DNA technology there is a compelling demand for tools able to perform efficient, exhaustive and integrative analyses of multiple microarray datasets. Specifically, what is particularly evident is the need to link the results obtained from these new tools with the wealth of clinical information. The final goal is to bridge the gap existing between biomedical researchers and pathologists or oncologists providing them with a common framework of interaction. To overcome such difficulty we have developed GENECBR, a freely available software tool that allows the use of combined techniques that can be applied to gene selection, clustering, knowledge extraction and prediction. In diagnostic mode, GENECBR employs a case-based reasoning model that incorporates a set of fuzzy prototypes for the retrieval of relevant genes, a growing cell structure network for the clustering of similar patients and a proportional weighted voting algorithm to provide an accurate diagnosis.

1 Introduction

In recent years, machine learning and data mining fields have found a successful application area in the field of DNA microarray technology. Gene expression profiles are composed of thousands of genes at the same time, representing complex relationships between them. One of the well known constraints specifically related to microarray data is the large number of genes in comparison with the small number of available experiments or cases.

Recent studies in human cancer have demonstrated that microarrays can be used to develop a new taxonomy of cancer, including major insights into the genesis, progression, prognosis, and response to therapy on the basis of gene expression profiles [1]. However, there continues to be a need to develop new approaches to (*i*) diagnose cancer early in its clinical course, (*ii*) more effectively treat advanced stage disease, (*iii*) better predict a tumors' response to therapy prior to the actual treatment, and (*iv*)

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 112–117, 2009.
springerlink.com

ultimately prevent disease from arising through chemopreventive strategies. Given the fact that systematic classification of tumor types is crucial to achieve advances in cancer treatment, several research studies have been developed in this direction [2, 3].

In this context, case-based reasoning (CBR) systems are particularly applicable to the problem domain because they (*i*) support a rich and evolvable representation of experiences/problems, solutions and feedback; (*ii*) provide efficient and flexible ways to retrieve these experiences; and (*iii*) apply analogical reasoning to solve new problems [4]. CBR systems can be used to propose new solutions or evaluate solutions to avoid potential problems. In the work of [5] it is suggested that analogical reasoning is particularly applicable to the biological domain, partly because biological systems are often homologous (rooted in evolution). Moreover, biologists often use a form of reasoning similar to CBR, where experiments are designed and performed based on the similarity between features of a new system and those of known systems.

In this sense, the work of [6] proposes a mixture of experts for case-based reasoning (MOE4CBR) developing a method that combines an ensemble of CBR classifiers with spectral clustering and logistic regression. This approach not only achieves higher prediction accuracy, but also leads to the selection of a subset of features that have meaningful relationships with their class labels. Previously, [7] showed their initial work in applying a CBR approach to the problem of gene-finding in mammalian DNA. The results obtained from their experiments indicate that it is certainly feasible to do DNA-to-DNA comparisons in order to isolate relevant coding regions. A previous successful work in the same area using CBR was carried out by [8].

In [9] the authors showed how their CASIMIR/CBR system was able to suggest solutions for breast cancer treatment by adapting the rules of a previous rule-based system (CASIMIR/RBR). The work of [4] demonstrated how case-based reasoning can be applied to assist in analyzing genomic sequences and determining the structure of proteins. They also provide an overview of several other applications in molecular biology that have benefited from case-based reasoning.

2 Software Architecture

GENECBR was implemented to support integrative work for interdisciplinary research groups working together in order to design, implement and test new techniques for supervised and unsupervised cancer classification and clustering. Figure 1 (left) shows this user-dependent architecture.

(1) *Pathologists or oncologists*: GENECBR (*diagnostic mode*) implements an effective and reliable system able to diagnose cancer subtypes based on the analysis of microarray data using a CBR architecture (see Figure 1 (right)).
(2) *Biomedical researches*: GENECBR (*expert mode*) offers a core workbench for designing and testing new techniques and experiments.
(3) *Programmers*: GENECBR (*programming mode*) includes an advanced edition module for run-time modification of previous coded techniques based on BeanShell (http://www.beanshell.org/).

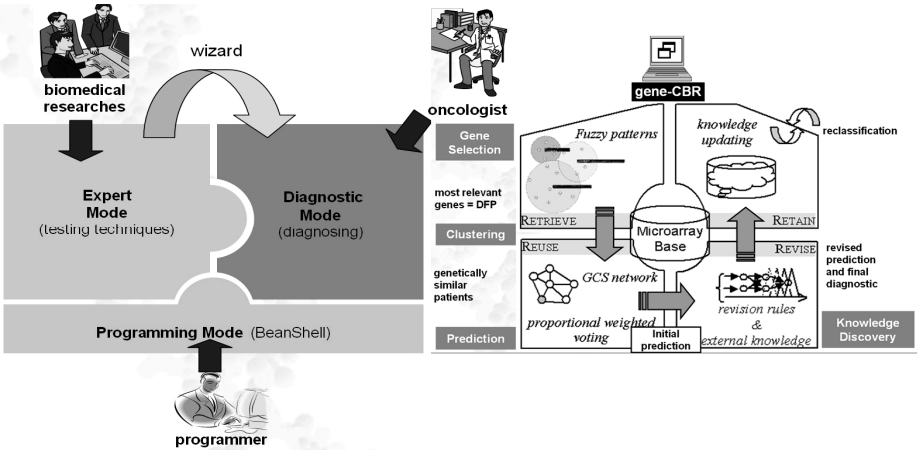


Fig. 1. (left) Logic architecture and interfaces of GENECBR system. (right) Life-cycle of GENECBR system working in *diagnostic mode.*

GENECBR is written entirely in Java 1.5 and portable across multiple operating systems and platforms. It is simple to install and easy to update through the utilization of Java Web Start technology (http://java.sun.com/products/javawebstart/) which ensures the execution of the last available version.

3 Core Functions and Features

GENECBR *expert mode* can load microarray expression datasets from any platform as input, as long as the data coming from different experiments have been summarized into a matrix of normalized expression values (microarray base). The input file should contain text fields with comma as separator and can hold lines starting with the number sign ('#') for commenting meta-data information. GENECBR includes a co-expression analysis module, a discriminant expression analysis module, a supervised/unsupervised classification module and various *add-ins* such us the net explorer module.

3.1 Co-expression Analysis Module

This module is used to identify sets of genes simultaneously co-expressed in multiple datasets belonging to patients suffering the same kind of cancer. Briefly, given a set of microarrays which are well classified, for each class a *fuzzy pattern* (FP) can be constructed from the *fuzzy microarray descriptor* (FMD) associated with each one of the microarrays [10]. The FMD is a comprehensible description for each gene in terms of one of the following linguistic labels: Low, Medium and High. Therefore, the fuzzy pattern is a prototype of the FMDs belonging to the same class where the membership criterion of each gene to the fuzzy pattern of the class is frequency-based. In this sense, we can assume that the genes included in a FP, are significant for the classification of any novel case within the class associated with that pattern.

3.2 Discriminant Expression Analysis Module

This module is developed to derive sets of genes expressed differentially in several FPs. The objective is to select those genes that allow us to discriminate the new case from one class with regard to the others. Here we introduce the notion of *discriminant fuzzy pattern* (DFP) with regard to a collection of fuzzy patterns. A DFP version of a FP only includes those genes that can serve to differentiate it from the rest of the patterns. The algorithm used to compute the DFP version of each FP in a collection of fuzzy patterns can be consulted in [11]. A consequence is that the computed DFP for a specific FP is different depending on which other FPs are compared with it. It's not surprising that the genes used to discern a specific class from others (by mean of its DFP) will be different if the set of rival classes also changes.

3.3 Supervised and Unsupervised Clustering Module

The DFP derived from the discriminant expression analysis module serves as a filter to select a reduced number of relevant and representative genes, allowing other artificial intelligence techniques to be able to tackle the high-dimensional data [12]. GENECBR incorporates a GCS neural network able to cluster all patients that are genetically similar given a selected group of genes and without taking into account their previous assigned classes. Our proposed method aims to find new relations between the patients as yet unknown. Therefore, it is possible and not contradictory to group together patients suffering different (but genetically related) diseases. Such a topology has the added advantage that inter-cluster distances can be precisely quantified. Since such networks contain explicit distance information, they can be used effectively to (*i*) represent an indexing structure which indexes sets of related patients and (*ii*) to serve as a similarity measurement between individual patients.

3.4 Net Explorer Module

The practice of biomedical research seeks to comprehend the complexity of complex organisms, or their subsystems, by combining many different kinds of data to improve existing knowledge. In current practice, as experts explore their data, they typically create manual, ad hoc connections among software tools and databases, cutting and pasting queries, creating temporary files, running web searches and taking notes. GENECBR includes an Internet explorer module able to gather additional information (gene annotations, public gene ids, biological functions, relevant related articles from PubMed/MedLine, etc.) in order to facilitate the integration of several sources of information. GENECBR always keeps their local databases updated, downloading new information as soon as it is available in Internet.

3.5 GENECBR Wizard

Every time the biomedical research group finishes their work GENECBR provides a guided 4-step wizard to setup GENECBR in diagnostic mode. With this configuration, the application is ready to receive a new sample (microarray experiment) and perform all the programmed tests in only one step.

4 Conclusions

We have presented a software package, GENECBR, as a successful application of a CBR system for cancer diagnosis using microarray datasets. GENECBR supports rich and evolvable representations of experiences, solutions and feedback. The experts often use a form of reasoning similar to CBR, whereby experiments are designated and performed based on the similarity between features of a new situation and those of known experiences. GENECBR allows the utilization of combined techniques that can be used for gene selection, clustering, knowledge extraction and prediction.

Using self-organizing GCS networks to meaningfully cluster filtered microarray data has a number of appealing advantages over other approaches such as incremental self-construction, and easy visualization of biological relationships between input data. The explanations of the clustering process carried out by the network can be addressed by means of our DFP vector. The most relevant knowledge for each cluster can be highlighted, and provide meaningful explanations about the clustering process and useful insight into the underlying problem and data.

The results obtained from the interrelationship of the proposed techniques are very promising and they back up the idea that CBR systems can offer a number of advantages as a framework for the microarray domain. Specifically, GENECBR allows us to obtain a more general knowledge about the problem domain and to gain a deeper insight into the importance of each gene related to each pathology.

Acknowledgements. We are deeply indebted to the personnel at the Hematology Service of the University Hospital of Salamanca and Cancer Research Center, who made this research work possible, by providing data and allowing us access to facilities.

References

1. Ochs, M.F., Godwin, A.K.: Microarrays in Cancer: Research and Applications. BioTechniques 34, s4–s15 (2003)
2. Xiang, Z.Y., Yang, Y., Ma, X., Ding, W.: Microarray expression profiling: Analysis and applications. Current Opinion in Drug Discovery & Development 6(3), 384–395 (2003)
3. Golub, T.: Genome-Wide Views of Cancer. The New England Journal of Medicine 344, 601–602 (2001)
4. Jurisica, I., Glasgow, J.: Applications of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special issue on Bioinformatics 25(1), 5–95 (2004)
5. Aaronson, J.S., Juergen, H., Overton, G.C.: Knowledge Discovery in GENBANK. In: Proc. of the First International Conference on Intelligent Systems for Molecular Biology, pp. 3–11 (1993)
6. Arshadi, N., Jurisica, I.: Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains. IEEE Transactions on Knowledge and Data Engineering 17(8), 1127–1137 (2005)
7. Costello, E., Wilson, D.C.: A Case-Based Approach to Gene Finding. In: Proc. of the Fifth International Conference on Case-Based Reasoning Workshop on CBR in the Health Sciences, pp. 19–28 (2003)

8. Shavlik, J.: Finding Genes by Case-Based Reasoning in the Presence of Noisy Case Boundaries. In: Proc. of the DARPA Workshop on Case-Based Reasoning, pp. 327–338 (1991)
9. Lieber, J., Bresson, B.: Case-Based Reasoning for Breast Cancer Treatment Decision Helping. In: Proc. of the 5th European Workshop on Case-Based Reasoning, pp. 173–185 (2000)
10. Fdez-Riverola, F., Díaz, F., Borrajo, M.L., Yáñez, J.C., Corchado, J.M.: Improving Gene Selection in Microarray Data Analysis using fuzzy Patterns inside a CBR System. In: Proc. of the 6th International Conference on Case-Based Reasoning, pp. 191–205 (2005)
11. Díaz, F., Fdez-Riverola, F., Glez-Peña, D., Corchado, J.M.: Using Fuzzy Patterns for Gene Selection and Data Reduction on Microarray Data. In: Proc. of the 7th International Conference on Intelligent Data Engineering and Automated Learning, pp. 1087–1094 (2006)
12. Díaz, F., Fdez-Riverola, F., Glez-Peña, D., Corchado, J.M., Applying, G.C.S.: Networks to Fuzzy Discretized Microarray Data for Tumour Diagnosis. In: Proc. of the 7th International Conference on Intelligent Data Engineering and Automated Learning, pp. 1095–1102 (2006)

Evolutionary Techniques for Hierarchical Clustering Applied to Microarray Data

José A. Castellanos-Garzón and Luis A. Miguel-Quintales

University of Salamanca, Department of Computer Science and Automatic, Faculty of Sciences, Plaza de los Caídos s/n, 37008 Salamanca, Spain
{jantonio,lamq}@usal.es
http://informatica.usal.es

Summary. In this paper we propose a novel hierarchical clustering method that uses a genetic algorithm based on mathematical proofs for the analysis of gene expression data, and show its effectiveness with regard to other clustering methods. The analysis of clusters with genetic algorithms has disclosed good results on biological data, and several studies have been carried out on the latter, although the majority of these researches have been focused on the partitional approach. On the other hand, the deterministic methods for hierarchical clustering generally converge to a local optimum. The method introduced here attempts to solve some of the problems faced by other hierarchical methods. The results of the experiments show that the method could be very effective in the cluster analysis on DNA microarray data.

Keywords: Hierarchical clustering, DNA microarray, evolutionary algorithm, genetic algorithm, cluster validity, combinatorial optimization.

1 Introduction

Genetic algorithms ([1, 2]) represent a powerful tool to solve complex problems where the traditional methods face difficulties in finding an optimal solution. The power of genetic algorithms (GAs) lies in its emulation of natural processes, such as adaptation, selection, reproduction and their merge with chance produces robust methods.

The application of GAs to *data mining* has significant importance in the knowledge extraction through the study of classification problems. Likewise, the analysis of clusters ([3, 4]) as an unsupervised classification is the process of classifying objects into subsets that have meaning in the context of a particular problem.

The hierarchical cluster analysis ([5, 6]) is a powerful solution for showing biological data using visual representations, where data can easily be interpreted through a spacial type of tree structures that show cluster relationships, such as the *dendrogram* representation. Overall, the clustering techniques applied to DNA microarray data ([7, 8]) have proven to be helpful in the understanding of the gene function, gene regulation, cellular processes, and subtypes of cells. The latter, it can be a tool very useful for the human health research.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 118–127, 2009.
springerlink.com

The first aspect of this research is that the problem of finding the best dendrogram on a data set, is a problem approached from the combinatorial optimization field, and it is an *NP-Complete* problem. Furthermore, it is not feasible to explore all possibilities on the dendrogram search space, and thereby arises the need of introducing methods that do not consider every solution in the search space, such as evolutionary techniques [9, 10].

The most important accomplishment of this work is the novel method definition of *hierarchical clustering based on GAs*, aimed at the search of global optimums into dendrogram search space on a data set. In addition, we present the theoretical basis of this method, as well as carry out a comparative analysis with other hierarchical clustering methods.

2 The Genetic Algorithm

In recent years, there has been an increasing interest in dealing with the problem of clustering using genetic algorithms, mainly for the partitional approach,[11, 12]. The method is aimed at the search of high quality clustering dendrograms.

The individuals (chromosomes) are dendrograms on a given data set, encoded as an ordered set of clusterings, where each clustering has an order number, called level. Initially, each dendrogram of a population is built up from an initial level to a higher level by joining two clusters chosen randomly in a level, in order to build the next one.

Length of an Individual

The length of a dendrogram can be defined as its number of levels (clusterings), but in the best case, until the half of the dendrogram levels, there will be unitary clusters[1] and that does not have a practical meaning, hence, those levels can be removed. Thus, a parameter can be introduced in order to remove the part of a dendrogram that does not give information. Therefore, we define the length of the dendrogram as follows:

Definition 1. *Dendrogram length.*
Set $\mathfrak{P}_n$ be a data set of size n and set $\mathfrak{G}$ be a dendrogram on $\mathfrak{P}_n$, then the length of $\mathfrak{G}$ is the clustering number of it and is defined as:

$$|\mathfrak{G}| = n - \lfloor n \cdot \delta \rfloor - 2, \tag{1}$$

where δ [2], is the part of $\mathfrak{G}$ to remove, assuming $\delta \geq 1/2$.

2.1 Fitness Function

In every GA it is necessary to measure the goodness of the candidate solutions. In this problem, the fitness of a dendrogram must be evaluated, hence we based

[1] One-element clusters.
[2] Is the fraction of $\mathfrak{G}$ that does not give information.

on one of the given definitions of cluster in [3], that is, *The objects inside of a cluster are very similar, whereas the objects located in distinct clusters are very different.* Thereby, the fitness function will be defined according to the concepts of both, *homogeneity* and *separation*, introduced in [13].

We begin by defining cluster homogeneity and afterwards defining more complex structures until reaching the dendrogram structure.

Definition 2. *Cluster homogeneity.*
If $\mathfrak{D} = [d(i,j)]$ is the proximity matrix on the $\mathfrak{P}_n$ data set, being d the defined metrics on this data set, $\mathfrak{C}$ a clustering of objects in $\mathfrak{P}_n$, C a cluster into $\mathfrak{C}$ and $m = |C|$, then the homogeneity of C is:

$$h(C) = \frac{2}{m \cdot (m-1)} \sum_{i \neq j}^{m \cdot (m-1)/2} d(i,j), (\forall i, j \in C). \tag{2}$$

Definition 3. *Clustering homogeneity.*
Set $\mathfrak{C}$ be a clustering of $\mathfrak{P}_n$, being $k = |\mathfrak{C}|$, then the homogeneity of $\mathfrak{C}$ is:

$$\mathcal{H}(\mathfrak{C}) = \frac{1}{k} \sum_{i=1}^{k} h(C_i). \tag{3}$$

Definition 4. *Distance between two clusters.*
Set $\mathfrak{C}$ be a clustering of $\mathfrak{P}_n$, set C_1 and C_2 be two clusters of $\mathfrak{C}$, then the distance d_m between these clusters is defined as:

$$d_m(C_1, C_2) = \min\{d(i,j)/i \in C_1, j \in C_2\}. \tag{4}$$

Definition 5. *Clustering separation.*
Set $\mathfrak{C}$ be a clustering of $\mathfrak{P}_n$, set C_1 and C_2 be two clusters of $\mathfrak{C}$, $k = |\mathfrak{C}|$, then the $\mathfrak{C}$ separation is:

$$\mathcal{S}(\mathfrak{C}) = \frac{2}{k \cdot (k-1)} \sum_{i \neq j}^{k \cdot (k-1)/2} d_m(C_i, C_j), \quad (\forall i, j \in [1, k]). \tag{5}$$

Definition 6. *Clustering fitness function.*
Set $\mathfrak{C}$ and $\mathfrak{D}$ be a clustering of objects in $\mathfrak{P}_n$ and the proximity matrix of $\mathfrak{P}_n$ respectively, then the fitness function of $\mathfrak{C}$ is defined as:

$$f_c(\mathfrak{C}) = \max \mathfrak{D} + \mathcal{S}(\mathfrak{C}) - \mathcal{H}(\mathfrak{C}). \tag{6}$$

Definition 7. *Dendrogram fitness function.*
Set $\mathfrak{G}$ and $\mathfrak{C}_i$ be a dendrogram on $\mathfrak{P}_n$ and a clustering of $\mathfrak{G}$ respectively, then the fitness function of $\mathfrak{G}$ is:

$$f_d(\mathfrak{G}) = \frac{1}{|\mathfrak{G}|} \sum_{i=1}^{|\mathfrak{G}|} f_c(\mathfrak{C}_i). \tag{7}$$

Based on the previous definition, an *ac agglomerative coefficient* can be used in order to estimate the level into a dendrogram $\mathfrak{G}$, where a cut can be carried out, that is:

Definition 8. *Agglomerative coefficient.*
Set $\mathfrak{G}$ and $\mathfrak{C}_i$ be a dendrogram on $\mathfrak{P}_n$ and a clustering of $\mathfrak{G}$, respectively. The agglomerative coefficient of $\mathfrak{G}$ is defined as:

$$ac(\mathfrak{G}) = \arg_{i \in [1,|\mathfrak{G}|]} \max f_c(\mathfrak{C}_i), \tag{8}$$

the level i whose clustering has the maximum fitness of the whole dendrogram.

2.2 Improving the Fitness Function Cost

Due to the computation complexity of the fitness function defined in (7), the need of decreasing its computation time has arisen. From the theoretical outlook, the above is verified in the following proposition:

Proposition 1. *Algorithmic complexity of f_c.*
Set $\mathfrak{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ be a clustering of a dendrogram $\mathfrak{G}$ on $\mathfrak{P}_n$ and set f_c be the fitness function defined in (6), then the order of $f_c(\mathfrak{C})$ is $O(k^2m^2)$ $(\Omega(kn^2))$, where $m = \max\{|\mathcal{C}_i|\}$, $i \in [1, k]$.

The fitness function defined in (7) can be transformed in an equivalent one but more efficient. This is shown in the following lemmas:

Lemma 1. *Recurrent homogeneity.*
Set $\mathfrak{C}_i$ be a clustering of level i and set $\mathfrak{C}_{i+1}$ be a clustering of level $i+1$, both in the dendrogram $\mathfrak{G}$; C_j and C_l, the two clusters of level i such that its join forms a new clustering $\mathfrak{C}_{i+1}$ of level $i+1$, then the homogeneity of $\mathfrak{C}_{i+1}$ $(\mathcal{H}_1(\mathfrak{C}_{i+1}))$ is computed in the following expression:

for $i = 1$, that is, for the first clustering, $\mathcal{H}_1(\mathfrak{C}_1) := \mathcal{H}(\mathfrak{C}_1)$ and for $i > 1$,

$$\begin{aligned}(k-1) \cdot \mathcal{H}_1(\mathfrak{C}_{i+1}) = k \cdot \mathcal{H}_1(\mathfrak{C}_i) - h(C_j) - h(C_l) + \tfrac{1}{k_3}[k_1 \cdot h(C_j) + \\ + k_2 \cdot h(C_l) + l_1 \cdot l_2 \cdot d_p(C_j, C_l)],\end{aligned} \tag{9}$$

where $k = |\mathfrak{C}_i|, l_1 = |C_j|, l_2 = |C_l|, k_1 = \binom{l_1}{2}, k_2 = \binom{l_2}{2}$, $k_3 = \binom{l_1 \cdot l_2}{2}$, and d_p is the average distance between the clusters C_j and C_l.

Lemma 2. *Recurrent separation.*
Keeping the same conditions of the previous lemma, one can obtain the recurrent separation of a clustering $\mathfrak{C}_{i+1}$ $(\mathcal{S}_1(\mathfrak{C}_{i+1}))$:

for $i = 1, S_1(\mathfrak{C}_1) := S(\mathfrak{C}_1)$ and for $i > 1$,

$$(g-k+1)\cdot S_1(\mathfrak{C}_{i+1}) = g\cdot S_1(\mathfrak{C}_i) - d_m(C_j,C_l) - \sum_{t\neq j\wedge t\neq l}^{k-2}[d_m(C_j,C_t) + \\ + d_m(C_l,C_t) - \min\{d_m(C_j,C_t), d_m(C_l,C_t)\}], \tag{10}$$

where $k = |\mathfrak{C}_i|, g = \binom{k}{2}$, being g the number of distances among the clusters of $\mathfrak{C}_i$.

Definition 9. *Clustering recurrent fitness.*
The fitness function of a clustering $\mathfrak{C}_{i+1}$ of $\mathfrak{G}$, according to $\mathcal{H}_1$ and S_1, is defined as:

$$g_c(\mathfrak{C}_{i+1}) = \max\mathfrak{D} + S_1(\mathfrak{C}_{i+1}) - \mathcal{H}_1(\mathfrak{C}_{i+1}), \tag{11}$$

known $\mathcal{H}(\mathfrak{C}_i)$ and $S(\mathfrak{C}_i)$.

Definition 10. *Dendrogram recurrent fitness.*
The fitness function of a dendrogram $\mathfrak{G}$, being $\mathfrak{C}_i$ a clustering of it is:

$$g_d(\mathfrak{G}) = \frac{1}{|\mathfrak{G}|-1}\sum_{i=1}^{|\mathfrak{G}|-1} g_c(\mathfrak{C}_i). \tag{12}$$

Once defined the recurrences, it is possible to verify that the cost of the fitness function defined in (11) is less than the cost of this one defined in (6).

Proposition 2. *Algorithmic complexity of g_c.*
Set $\mathfrak{C}_i, \mathfrak{C}_{i+1}$ be two clusterings (levels i and $i+1$) of a dendrogram $\mathfrak{G}$ on $\mathfrak{P}_n$, $k = |\mathfrak{C}_i|$ and $m = \max\{|C_j|\}$, $j \in [1,k]$, then the temporal complexity of $g_c(\mathfrak{C}_{i+1})$ is $O(km^2)$ $(\Omega(n^2))$.

2.3 Mutation Operator

The mutation of a dendrogram is performed according to the following steps:

1. We will consider two parameters τ and ϵ for each dendrogram $\mathfrak{G}$ where:
   - τ is the percentage of choosing cluster pairs into level i to build the following level $i+1$;
   - ϵ is a small value that represents the similarity between two clusters, according to the homogeneity measure.
2. A random number $i \in [1, |\mathfrak{G}|]$ is generated, it is the level where the mutation of $\mathfrak{G}$ is carried out.
3. For the clustering of the level i of the previous step, one of the following conditions is chosen:

- the most homogeneous join of cluster pairs of $\tau\%$ of random cluster pairs is chosen;
- the cluster pair with a difference from the cluster pair chosen in the above condition less or equal than ϵ is chosen.

4. The cluster pair chosen in the previous step is joined in order to form a new cluster so that the clustering of the next level $i+1$ can be built.
5. The steps 3 and 4 are repeated on the new level, until i reaches the level $|\mathfrak{G}|$.

2.4 Crossover Operator

The crossover is carried out on two dendrograms to obtain a child dendrogram, and is based on the idea of [14], that is:

1. Given two dendrograms $\mathfrak{G}_1$ and $\mathfrak{G}_2$ (parents), a random number i in $[1, |\mathfrak{G}_1|]$ is generated to choose the level where one can carry out the crossover between both dendrograms.
2. Through a strategy of *greedy algorithm*, the best $\lfloor k/2 \rfloor$ clusters[3] of level i of both dendrograms of the above step are chosen, being k the number of clusters of the level i. A new clustering is formed by repairing the chosen clusters [14, 15].
3. As soon as the new clustering for the level i is built, one can build up the new dendrogram:
   - the higher levels to the level i are built using the MO;
   - the lower levels to the level i are built in a divisible way, that is, for each level less than i, the less homogenous cluster is chosen to be split in two; This process is repeated until reaching the first level.
4. The parent of the less fitness value is replaced by the child dendrogram of the step 3.

3 Experiments on Gene Expression Data

In this section we study the behavior of the GA on a simulated data set of gene expression data and compare the results with other methods according to some cluster validity measures [16, 13, 4]. This data set is considered as benchmark data to prove different clustering algorithms and it was used in [17], published in `http://faculty.washington.edu/kayee/cluster`. The expression matrix of this one is composed of 384 genes evaluated on 17 conditions, labeled into 5 clusters of genes (*ground truth*) and it was normalized with mean 0 and variance 1. The method was implemented on the *R language* (R Development Core Team, [18]).

3.1 Goodness of the Individuals

In this subsection, the curves described by the fitness values of the clusterings in a dendrogram, of the initial and final population, are shown using graphs,

[3] The most homogenous clusters of both dendrograms.

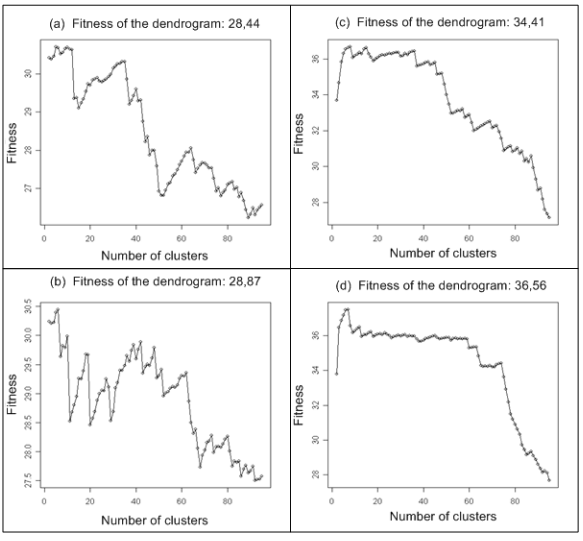


Fig. 1. Graphs of four dendrograms (from a to d), a and b belong to the initial population, while c and d are in the last population. The fitness values of each clustering are shown for each dendrogram.

where the cluster number ($x - axis$) of each clustering of a dendrogram vs. the fitness value ($y - axis$) of each of its clustering is plotted. The graphs of two dendrograms in the initial population (a and b) and two dendrograms of the final population (c and d) are shown in figure 1, $\delta = 3/4$. As it is shown in this figure, the curve described by the dendograms of the initial population presents many oscillations and there are many large differences between the fitness values of two consecutive clusterings. However, the above individuals are improved by execution of the GA with the following parameters: the values of the crossover and mutation likelihood were assigned around 60% and 5% respectively, the generation number in $[10^3, 10^6]$, $\tau \in [15\%, 40\%]$, $\epsilon = 3\%$, $x = 90\%$ and the *euclidian distance* was used on the data set.

One can observe that the fitness of the individuals in the last generation (c and d) was improved after executing of the GA. Hence, the problems presented on the individuals of the initial population have been reduced after applying the genetic operators.

3.2 Homogeneity, Separation and Agreement with the Reference Partition

In this subsection we are going to carry out a cluster validity process to compare the results of the GA. Therefore, we have focused on the quality of clusters in terms of homogeneity (Homog), separation (Separ) [13], silhouette width (SilhoW, [4]), the *Jaccard coefficient* (JC), and the *Minkowski measure* (MM) [13],

Table 1. Cluster validity of the GA vs. five hierarchical clustering methods

Method	f_d	Cluster (ac)	f_c	Homog	Separ	SilhoW	JC	MM
Ground truth	-	5	31.10	6.13	6.60	36.54	-	-
Agnes	32.03	30	37.50	2.02	7.64	36.80	0.12	1.20
Diana	35.41	9	37.47	2.76	9.07	36.96	0.16	1.40
Eisen	23.54	3	39.74	5.80	22.63	37.29	**0.23**	1.82
HybridHclust	37.96	52	39.68	**1.45**	6.72	36.72	0.06	**1.01**
Tsvq	37.32	**5**	40.05	3.86	12.57	37.16	0.15	1.37
GA:								
1.-	32.68	22	35.28	6.82	6.44	36.13	0.09	**1.12**
2.-	33.71	8	36.68	6.70	5.20	36.15	0.21	1.73
3.-	34.41	7	37.49	6.59	5.28	36.09	**0.23**	1.79
4.-	36.56	3	**44.14**	6.36	10.37	36.40	**0.23**	1.81
5.-	37.13	2	39.34	6.27	**28.87**	**37.36**	**0.23**	1.83
6.-	37.55	4	43.11	6.16	11.57	36.32	0.21	1.79
7.-	37.74	3	43.40	6.28	17.02	36.33	**0.23**	1.81
8.-	**39.20**	4	43.51	6.15	15.99	36.29	0.21	1.80

for five methods of hierarchical clustering with *mean link* as a type of distance; *Agnes* and *Diana*[4], *Eisen*[5], *HybridHclust* [19] and *Tsvq*[20].

The GA was initialized with a population of 10 individuals, δ in $\{3/4, 4/5, 5/6, 12/13\}$ and the other parameters were assigned as in the above section. The best 8 outputs were extracted to make comparisons, such as listed in table 1, where the GA is compared with five methods according to eight measures: the *Cluster* column is the number of the cluster of the best clustering in a dendrogram (using ac coefficient) then, for that same clustering the other measures located in the right side of the *Cluster* column of the table were computed. The *Ground truth* row contains the evaluations on the 5 pre-classified clusters of the data set and the best values of the method-measure are highlighted in that table.

In table 1 it can emphasize on different results referring to the GA, where the convergence is proven in the f_d column, since the fitness values of the solutions can be improved. Furthermore, the values of the *Cluster* column, could be employed to determine the optimal number of the cluster by applying some statistical indicator on this list of values. Due to the above results, the method reached the best values for f_d, f_c, separation and silhouette width indicator. For the MM and JC coefficient, it can be emphasized that four executions of the GA and the *Eisen* method reached the best results on JC. In contrast, one of the executions of the GA and the *HybridHclust* method reached the best results on MM.

4 Conclusion

The main goal of this paper has been to present and discuss the theoretical results of a novel evolutionary approach for hierarchical clustering, leading to

the search of global optimums in the dendrogram space. In order to show the effectiveness of this approach with regards to other methods, we have used a simulated data set of gene expression, published as a benchmark.

The introduced method achieved good experiments results in relation to other methods on the DNA microarray data. Therefore, this method can be very important in the process of knowledge discovery as well as in the analysis of gene expression data. Moreover, the most natural way of genetic algorithm application lies precisely in the study of biological processes. Finally, the most important outcomes are:

1. Two fundamental lemmas for improving the temporal complexity of the fitness function. Moreover, the complexity of any other fitness function can be reduced, based on the proof given in those lemmas;
2. The flexibility of the GA to change the genetic operators or add other heuristics, is a strong tool for clustering;
3. The method performed well respect to both, the definition of clustering, that is, homogeneity and separation; and a reference partition of the chosen gene expression data set.

References

1. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley Longman, Inc. (1989)
2. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 3rd edn. Springer, Heidelberg (1999)
3. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1998)
4. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. An Introduction to Clustering Analysis. John Wiley & Sons, Inc., Hoboken (2005)
5. Eisen, M., Spellman, T., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, USA 95, 14863–14868 (1998)
6. Jiang, D., Pei, J., Zhang, A.: DHC: A density-based hierarchical clustering method for time series gene expression data. In: Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE) (2003)
7. Berrar, D.P., Dubitzky, W., Granzow, M.: A Practical Approach to Microarray Data Analysis. Kluwer Academic Publishers, New York (2003)
8. Speed, T.: Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC Press LLC (2003)
9. De-Jong, K.A., Spears, W.M.: Using Genetic Algorithms to Solve NP-Complete Problems. In: Proceedings of the Third International Conference on Genetic Algorithms (1989)
10. Godefriud, P., Khurshid, S.: Exploring very large state spaces using genetic algorithms. In: Katoen, J.-P., Stevens, P. (eds.) TACAS 2002. LNCS, vol. 2280, pp. 266–280. Springer, Heidelberg (2002)
11. Chu, P.C., Beasley, J.E.: A genetic algorithm for the set partitioning problem. Technical report, Imperial College, The Management School, London, England, 481–487 (1995)

12. Maulik, U., Bandyopadhyay, S.: Genetic algorithms-based clustering technique. The Journal of the Pattern Recognition Society 33, 1455–1465 (2000)
13. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering 16(11), 1370–1386 (2004)
14. Greene, W.A.: Unsupervised hierarchical clustering via a genetic algorithm. In: IEEE Congress on Evolutionary Computation, CEC 2003, vol. 2, pp. 998–1005 (2003)
15. Castellanos-Garzón, J.A., Miguel-Quintales, L.A.: Algoritmos genáticos para clustering de datos de expresión génica. Master's thesis, Computer Science and Automatic Department, University of Salamanca, Spain (2006)
16. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis 21, 3201–3212 (2005)
17. Yee-Yeung, K.: Clustering Analysis of Gene Expression Data. PhD thesis, University of Washintong (2001)
18. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2006) ISBN 3-900051-07-0
19. Chipman, H., Tibshirani, R.: Hybrid hierarchical clustering with applications to microarray data. Biostatistics 7, 302–317 (2006)
20. Macnaughton-Smith, P., Williams, W.T., Dale, M.B., Mockett, L.G.: Dissimilarity analysis: a new technique of hierarchical subdivision. Nature 202, 1034–1035 (1965)

Beds and Bits: The Challenge of Translational Bioinformatics

Daniel Glez-Peña[1], Pablo Vicente Carrera[2], Gonzalo Gómez López[2,3], and Carmen M. Redondo Marey[2]

[1] ESEI: Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
dgpena@uvigo.es
[2] Biomedical Foundation of Complejo Hospitalario Universitario of Vigo, Hospital do Rebullón, 36200, Vigo, Spain
credondo@fundacionbiomedica.org
[3] Spanish National Cancer Research Centre, 3rd Melchor Fernández Almagro St., 28029, Madrid, Spain
ggomez@cnio.es

Abstract. Since the Biomedical Foundation Bioinformatics group started in 2004, scientific activities, collaborations and staff have been continuously increasing. The Bioinformatics group develops its own scientific projects focused in translational Bioinformatics, joining so the biological observations to the clinical context. Hence, the main topics currently covered by our group are microarray data analysis, functional analysis and protein studies (structure, function and dynamics). Moreover, the Biomedical Foundation works with other scientific and academic institutions such as INB-CNIO, IIB and Vigo University. These enriching collaborations provide of training, new projects and resources to our group. For instance and, by means of an agreement with the INB, the Foundation has displaced a postdoctoral fellow at CNIO Bioinformatics Unit. In this paper, we report the current Bioinformatics activities carried out by the Biomedical Foundation Bioinformatics group.

1 Introduction

The Biomedical Foundation of Vigo University College Hospital (CHUVI, www.fundacionbiomedica.org) manages, promotes and supports the clinical and basic research of public hospitals in the Galician city of Vigo. One of the main goals since the establishment of the Foundation has been to integrate new technologies and experimental techniques into the clinical framework in order to provide the best possible resources to patients, clinicians and researchers in their daily activity at hospitals. In this sense, the Foundation wants to stimulate research groups to translate their new biological observations into the clinical context. Thus, keeping in mind the applicability of the research results into the clinical environment, the Foundation's Bioinformatics group was established in 2004 with an initial staff constituted by 1 head scientist and 2 predoctoral fellows. This initial group focused their research on data and functional analysis of genomic high throughput (HT) studies.

At present, the CHUVI´s Bioinformatics group includes 2 postdocs, 1 predoctoral fellow and 1 student. The group has added more research lines including protein

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 128–136, 2009.
springerlink.com

studies and it will be included in a bigger Bioinformatics group constituted by the association of CHUVI´s Foundation and the SING group (http://sing.ei.uvigo.es/~riverola/) from the University of Vigo. The agreement between Vigo University and the Biomedical Foundation will produce a Bioinformatics Support Unit providing more human and technical resources to the project.

Despite of the fact that the Foundation is physically located in the hospitals of Vigo (Meixoeiro, Xeral-Cíes, Nicolás Peña and Rebullón) the research activity of the group takes place in Ourense and Madrid as well. From the very beginning the Foundation has supported and facilitated the training and mobility of researchers. In this way, due to the collaboration agreement signed by the Foundation and the Spanish Bioinformatics Institute (INB, www.inab.org), the Bioinformatics group have displaced a postdoctoral fellow (since June 2006) to the Bionformatics Unit (UBio, http://bioinfo.cnio.es) at Spanish National Cancer Research Institute (CNIO, www.cnio.es). This Unit is an active part of the CNIO´s Structural Biology and Biocomputational program (headed by Prof. Alfonso Valencia) and it is focused on the problems related to biological data management (integration, visualization, and knowledge extraction) with an emphasis on addressing specific user needs in relation to HT experimental methodologies. In collaboration with CNIO experimental research groups, external groups, and other technical units, CNIO´s Bioinformatics group assesses and adapts public and commercial data analysis tools and methods, and develops new ones when required. Moreover, the Unit is responsible for maintaining the scientific computational infrastructure of the CNIO.

In this INB/CNIO-Biomedical Foundation-Vigo University context, we daily work and collaborate with several experimental research groups on projects related to data analysis, evaluation of the currently used tools, analysis of large-scale data sets and retrieval and analysis of medical information from publicly available databases. By this reason, to get in touch with new HT methods, protocols and experimental technologies, it is essential to allow for a fluent communication between wet lab groups and Bioinformaticians. Furthermore, the translational scenario in which the Biomedical Foundation and many CNIO groups are focused demands clinical knowledge (pathologies, therapies, patients' management, etc.) in order to encompass biological observations and clinical applications. Here, translational Bioinformatics appears in order to deal with the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, genomic data in particular, into proactive, predictive, preventive, and participatory health.

This report describes in brief the currently work lines developed in our group and, also, those carried out as result of the different collaborations. INB/CNIO- Biomedical Foundation will be extensively described. Tools available in the group, techniques and analyses are described as well.

2 Research Lines

2.1 Microarray Data Analysis

Based on the experience obtained in collaborations with wet-lab groups, the CNIO Unit has focused its efforts on integrating tools and methods for the functional interpretation of the results of experimental datasets, especially expression data.

On the one hand, and in collaboration with the CNIO´s Lymphoma Group (M.A. Piris), several clustering methods have been applied to classify Gene Expression Profiling for almost 200 cases of B-cell Non-Hodgkin lymphomas, including Diffuse Large B-Cell Lymphomas (DLBCL), Mucosa-Associated Lymphoid Tissue lymphoma (MALT), Mantle Cell Lymphomas (MCL), Chronic Lymphocytic Leukemia(CLL), Follicular Lymphomas(FL), Marginal Zone Lymphoma (MZL), Spleenic Marginal Zone Lymphoma (SMZL) and Burkitt Lymphomas (BL). The functional signatures that were identified as distinguishing between these lymphoma types were defining cell cycle, apoptosis, cytokine-cytokine receptor interaction, T-cell receptor, B-cell receptor, cell adhesion, NF-kB activation, and other significant interactions. Moreover these signatures reveal sub-classes for diagnosed lymphoma types, suggesting the existence of a distinct functional heterogeneity among CLL, MCL and DLBCL. Comparison between these lymphoma clusters, following this definition, gave as a result the finding of a large number of genes which are different on each of these clusters; this list includes already known genes and a number of new potential markers and therapeutic targets.

On the other hand, clustering methods are been employed to classify epigenetic data as well. In collaboration with CNIO´s Cancer Epigenetics Group (M. Esteller), the Unit is building clustering maps for DNA methylation patterns observed in HBV and HPV virus. Within this goal, several different linkage methods (Average, Single, Complete…) and distances (Euclidean and Manhattan distances, Pearson and Spearman correlations…) are being applied to classify and visualize samples based on their methylation patterns.

In addition, the CNIO´s Bioinformatics Unit has been working on the application of techniques such as Gene Set Enrichment Analysis (GSEA)[1], which is able to use information on blocks of functionally related genes in order to build and test functional hypotheses on top of previously acquired biological information (Fig. 1 shows an example). Other collaborations with CNIO´s experimental groups are related to technical support for researchers such as data normalization, statistical support, functional analysis of gene lists, training in data analysis for CNIO staff and so on. Furthermore, the Unit collaborates actively with biotechnology and pharmaceutical private companies such as BioAlma and Pharmamar.

In the development area, CNIO´s Bioinformatics Unit is participating in the Cancer And Related Genes Online (CARGO) [2] server project. CARGO is a configurable biological web portal designed as a tool to facilitate, integrate and visualize results from biological internet resources, independently of their native format or access method. Through the use of small agents, called widgets, CARGO provides pieces of minimal, relevant and descriptive biological information. The final goal of CARGO is helping users in functional analysis of gene lists providing an intuitive and friendly interface integrating scattered biological information from different sources. The server is publicly available at http://cargo.bioinfo.cnio.es.

Concerning non-CNIO collaborations, CHUVI's Biomedical Foundation and Childhood Tumors Group Instituto de Investigaciones "Alberto Sols" (IIB, www.iib.uam.es) have signed a collaboration agreement providing bioinformatics support for genomics HT studies carried out in neuroblastomas and Ewing´s sarcomas. The collaboration has been centred in the Cyclic Loess method implementation for normalization of CodeLink

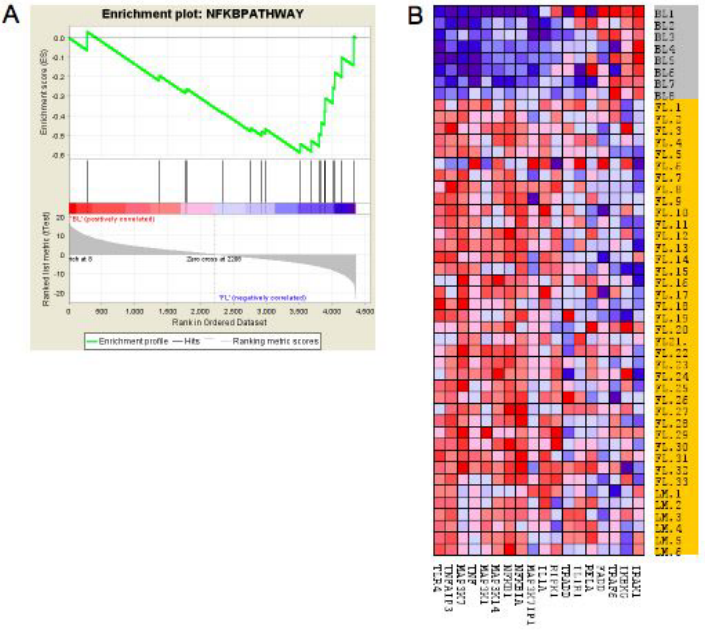


Fig. 1. Gene Set Enrichment Analysis of NFkB pathway in Follicular (FL) and Burkitt (BL) lymphomas using Biocarta annotations. A. Applying Kolmogorov-Smirnov test to find asymmetries in expression levels of functionally related blocks of genes, NFkB pathway genes were found as enriched in FL in comparison to BL. B. Heat map for expression values of the genes involved in NFkB pathway.

microarray data [3]. Differential expression analysis using several statistical approaches (parametric and non-parametric methods) has been carried out for gene selection as well. However, the collaboration core has been the functional interpretation of the data applying GSEA and commercial software for data mining. This approach has been also employed in other projects of the Childhood Tumors Group directed to the analysis of the DAX1 oncogene activity. As results of this collaboration two articles has been submitted to Oncogene (which is currently under revision) and Cancer Research journals respectively.

Other Biomedical Foundation project is related to the search for molecular markers useful in operating theatre. On hospital scenario in Vigo, we have employed Affymetrix microarrays in order to analyze the expression of 8500 known genes in lung tissues from 11 adenocarcinoma patients. Samples were taken at operating theatre from 10 healthy tissues obtained more than two centimetres from tumor limits (H), 10 non-pathological tissues in tumour limits (P) and 9 tumoral tissues (T). After differential expression analysis and multiple testing correction 24 statistically significant genes were found. Hierarchical clustering (see Figure 2) and PCA analysis, from 24 genes selected, show two groups clearly differentiated between H and T-arrays, 8 out of 10 P-arrays were included into the H-cluster. This work abstract was accepted in II IASLC International Workshop on Early Invasive Lung Cancer and in the European Conference on Computational Biology (ECCB07).

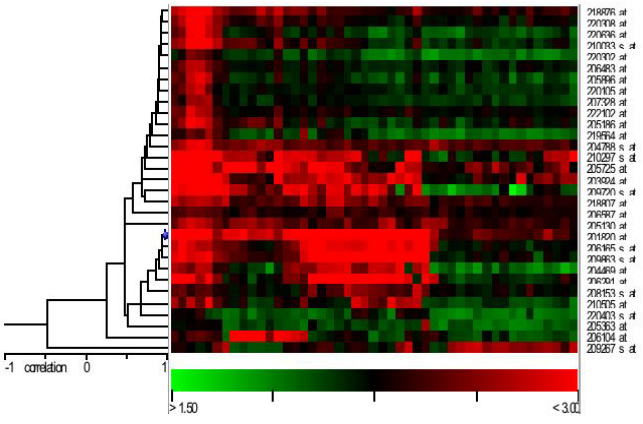


Fig. 2. UPGMA for 24 differentially expressed genes among H, P and T samples of lung adenocarcinoma. Selected genes were capable to clearly discriminate between H and T samples. Two out 10 P samples were clustered in the T subset. These samples showed tumoral cells by histological analysis.

2.2 Protein Studies

At the beginning of this year 2007 the arrival to the Biomedical Foundation of a new postodc with experience on Quantum Time-Dependent Dynamics of complex biological systems provoked the opening of a new research line in the Biomedical Foundation Bioinformatics group. Coming out of this new research line the group has now two new on-going research projects both of which include the collaboration with different clinicians of Vigo´s hospitals and with Bioinformaticians of the University of Vigo SING group.

The first project is focused on the finding of a protein pattern related with different cancer types (breast, lung and colon). The idea behind this project is to identify a group of proteins which can be of use on the detection, the prognosis and the following of the different cancer types.

The first step of this project is to retrieve all the information already available at public databases (www.ncbi.nlm.nih.gov/entrez) using for this work a search tool, BioSearch, developed by the SING group. This tool is allowing us to retrieve, classify, export and visualize big amounts of biomedical data. The Biomedical Foundation helps towards the continuous development of this tool by detecting needs and suggesting ideas for its improvement. The obtained information is added to some other obtained from the Human Protein Reference Database (www.hprd.org), the Human Proteinpedia (www.humanproteinpedia.org), the Human Protein Atlas (www.proteinatlas.org), the Universal Protein Resource database (www.uniprot.org) and the Genecards database (www.genecards.org). These databases provide information not only about the studied protein but also about the associated genes and related illnesses.

Following the finding of these huge amounts of data we classify the information according to several scientific criteria: clinical ones (metastasis, 5 years survival, 3 years disease-free), informatics criteria (information´s quality and consistency) and

biological criteria (experimental methodology used for protein determination, protein solubility in plasma/serum, etc.). These criteria allow us for the obtaining of a first protein pattern associated with each type of cancer.

The third step of the project is to complete this information with the one coming from the metabolic pathways available at KEGG database (www.genome.jp/keggpathway.html); for this step we will be using Pathway Architect (commercial software program available through CNIO´s collaboration) and other in-house developed tool for pathway comparison (available through SING group´s collaboration). The final set of proteins determined for each type of cancer will be thoroughly studied both by experimental and computational tools. From the Bioinformatics point of view protein structure and dynamics will focus our attention.

The ultimate goal of this ambitious project is to translate the experimental/theoretical information obtained through this study to the clinical world so these proteins could be used for cancer early detection, prognosis and following, facilitating so the clinicians´ work and, most importantly, improving the survival rate of the affected patients.

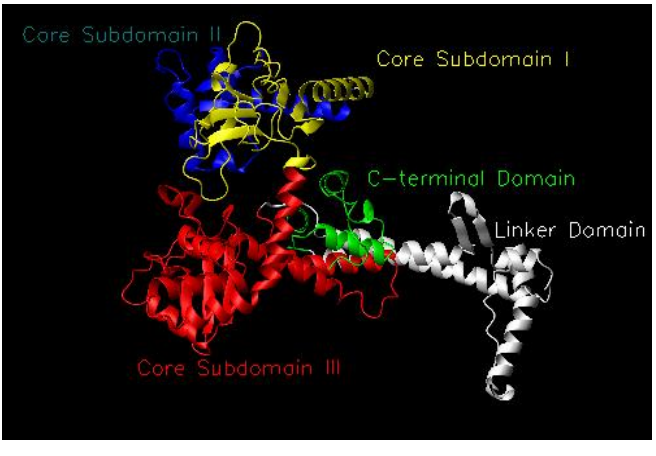


Fig. 3. Predicted three dimmensional structure of DNA-topoisomerase I as it was determined by using the Swiss-Pdbv program. The protein different domains where also determined.

A second project which is, at the moment, focusing our attention is related with the HIV. A group of clinicians is carrying out a study to determine different haplotypes found in the HIV infected patients. This study will allow them to classify the patients as progressors or long term non-progressors by using genetic criteria (at the moment this classification is made by following the patients´ evolution through the time). It will also allow determining which patients can have adverse reactions to some kinds of antiretroviral medicines avoiding then its prescription in these cases. The Biomedical Foundation Bioinformatics group will be responsible for analysing the obtained experimental data extracting the maximum knowledge from the available data.

3 Tools

3.1 Genomic High throughput Studies

Working on collaboration has several benefits on day-by-day tasks including the accessibility to different microarray platforms and a choice of tools for dealing with large sets of expression data and genomic functional analysis. Consequently, the group is working with data coming from spotted CNIO´s Oncochip v2.0 and commercial platforms (Agilent, GE CodeLink and Affymetrix).

In order to deal with this data, we use publicly available tools and commercial software as well. Thus, R (www.r-project.org) and Bioconductor libraries (www.bioconductor.org) are commonly employed to set up normalization methods and analyses protocols. GSEA and FatiScan [4] are employed for functional analysis of gene sets as mentioned above. The Connectivity Map [5] is applied to relate molecular signatures to drug treatment. Asterias (http://asterias.bioinfo.cnio.es), GEPAS (http://gepas.bioinfo.cipf.es) and Babelomics (http://babelomics.bioinfo.cipf.es) public servers for data preprocessing, filtering, annotation tasks, etc.

Regarding commercial software, the CNIO Unit has recently purchase an Ingenuity Systems IPA license (www.ingenuity.com). This tool is basically employed in gene list functional analysis, protein interactions network modelling, gene annotation, array data interpretation, etc. Other tools implementing exploratory techniques have also been used, including PathwayArchitect (Stratagene) for relevance network modelling, GeneSpring MX (Agilent), AKS2 (BioAlma), SPSS, etc. In addition, Biomedical Foundation provides and Acuity 4.0 license (Molecular Devices) for normalization, exploratory analysis, supervised and non-supervsed clustering, differential gene expression analysis, etc.

3.2 Protein Studies Tools

In order to analyse the HIV experimental data both, commercial and publicly available tools, will be used (the R project, already mentioned above, and the SPSS package, http://www.spss.com).

Some of the tools used for the protein studies have already being mentioned in Section 2.2. The protein structure will be firstly determined by using Swiss-pdb viewer (http://expasy.org/spdbv) which allow for an homology structure determination by comparing the studied protein with a similar one with an already X-ray obtained structure (available as pdb file); the refined protein structure and the dynamics of it will be determined by using NAMD (http://www.ks.uiuc.edu/Research/namd), a parallel-designed program for dynamics calculations of big molecular systems. The dynamics of the residues involved in the reaction will be determined by Quantum Mechanical Calculations and implemented by using a system-bath approach developed by one of the members of the Biomedical Foundation Bioinformatics group [6].

4 Conclusions

The Biomedical Foundation Bioinformatics Group started its scientific activities in 2004. Since the very beginning we have focused our efforts in translational bioinformatics

trying to bridge basic research to the clinical context. Following this rationale, the bioinformatics group develops its own scientific and academic projects. Hence, the main topics currently covered by our group are microarray data analysis, functional analysis and protein studies (structure, function and dynamics).

As explained above, the Biomedical Foundation Bioinformatics Group is also collaborating with other scientific and academic institutions. The INB-CNIO collaboration is specially enriching allowing us to collaborate with experimental labs located in a cancer research centre of excellence. At the moment, expression microarray data analysis of a large collection of B-cell lymphomas, functional analysis of cytogenetics of Ewing´s sarcoma and clustering of tumoral epigenetics in hepatocarcinomas are being carried out thanks to this agreement. Additionally, the human resources and infrastructure of our group will be improved by means of the commitment with Vigo University.

These enriching agreements provides training, new projects and resources to our group, and bridge the gap between wet labs needs and computational biology solutions available. In this way, Biomedical Foundation is setting up molecular biology and experimental surgery laboratories at Rebullón hospital. These new labs should supply new information, techniques and experimental validation to computational predictions helping in translation to routinely clinical protocols at hospital.

However, introducing bioinformatics into hospitals is not an easy task yet. As proposed in the main title of this report joining beds and bits is not trivial. Biocomputational technologies are, in most of the cases, completely new for clinicians and hospital staff. Biomedical Foundation proposal in bioinformatics should help in familiarizing hospital employees with this new knowledge area and its promising possibilities. Our aim in a medium term is to get clinicians used to bioinformatics and computational biology in order to take this tools and technologies into account in the clinical projects design.

Acknowledgments

David G. Pisano for his advice, support and confidence in the daily work at CNIO Bioinformatics Unit. Enrique Caso for his open-minded point of view in this collaboration.

References

1. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., Mesirov, J.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102(43), 15545–15550 (2005)
2. Cases, I., Pisano, D., Andres, E., Carro, A., Fernandez, J., Gomez-Lopez, G., Rodriguez, J., Vera, J., Valencia, A., Rojas, A.M.: CARGO: a web portal to integrate customized biological information. Nucleic Acids Res. 35(Web Server issue), W16–W20 (2007)
3. Wu, W., Dave, N., Tseng, G.C., Richards, T., Xing, E.P., Kaminski, N.: Comparison of normalization methods for CodeLink Bioarray data. BMC Bioinformatics 6, 309 (2005)

4. Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., Dopazo, J.: From genes to functional classes in the study of biological systems. BMC Bioinformatics 8, 114 (2007)
5. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R.: The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313(5795), 1929–1935 (2006)
6. Redondo, C.M.: Ph.D. Thesis. Femtodynamics of double proton transfer reactions. University College London, Londres, UK (2002)

A Matrix Factorization Classifier for Knowledge-Based Microarray Analysis

R. Schachtner[1], D. Lutter[2], A.M. Tomé[3], G. Schmitz[4], P. Gómez Vilda[5], and E.W. Lang[1]

[1] CIML/Biophysics, University of Regensburg, D-93040 Regensburg, Germany
elmar.lang@biologie.uni-regensburg.de
[2] CMB/IBI, GSF Munich, Germany
[3] IEETA/DETI, Universidade de Aveiro, 3810-193 Aveiro, Portugal
[4] Clinical Chemistry, University Hospital Regensburg, D-93042 Regensburg, Germany
[5] P. Gómez Vilda, DATSI/FI, Universidad Politécnica de Madrid, E-18500 Madrid, Spain

Summary. In this study we analyze microarray data sets which monitor the gene expression levels of human peripheral blood cells during differentiation from monocytes to macrophages. We show that matrix decomposition techniques are able to identify relevant signatures in the deduced matrices and extract marker genes from these gene expression profiles. With these marker genes corresponding test data sets can then easily be classified into related diagnostic categories. The latter correspond to either monocytes vs macrophages or healthy vs Niemann Pick C diseased patients. Our results demonstrate that these methods are able to identify suitable marker genes which can be used to classify the type of cell lines investigated.

1 Introduction

High-throughput genome-wide measurements like microarrays are traditionally analyzed applying two strategies: a) supervised approaches like support vector machines (SVM) and b) unsupervised approaches like projective subspace techniques. Any statistical analysis of gene expression probe level data, however, has to face the "large N, small M" problem setting [4], where N denotes the number of genes (= features, variables, parameters) and M denotes the number of samples (= experiments, environments, tissues). Also over-fitting has to be avoided to construct a classifier with a good generalization ability ([9]). Any robust classifier needs a sample-per-feature (SpF) ratio of 5 - 10-fold, while with usual microarray probe level measurements the SpF amounts to $1/50 - 1/200$ roughly. Most classical statistical tools fail in such "large N, small M" problem settings. Hence a substantial reduction of the feature space dimensionality via gene or feature selection is often the only way out of this sample-per-feature dilemma.

In this study we propose to include diagnostic knowledge and explore the potential of unsupervised techniques like independent component analysis (ICA) and nonnegative matrix factorization (NMF) to identify and extract marker

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 137–146, 2009.
springerlink.com

genes from microarray data sets and classify these data sets according to the diagnostic classes they represent. We focus on a recently proposed combination of such feature generation methods with diagnostic knowledge to build an appropriate classifier [8]. The feature generation methods expand the data in a new basis with a strongly reduced dimensionality. The projections of the data vectors onto the new basis vectors are called *features*. Alternatively one can categorize such methods, the most prominent among them are principal component analysis (PCA), ICA or NMF, as matrix decomposition techniques which extract *eigenarrays* (PCA), *expression modes* (ICA) or *metagenes* (NMF). These matrix decomposition approaches can be used also to build classifiers which allow to classify gene expression signatures or profiles into different classes [1].

Here we used M-CSF dependent in vitro differentiation of human monocytes to macrophages as a model process to demonstrate the potential of the proposed semi-supervised approach to build diagnostic classifiers, extract biomarkers and identify complex regulatory networks.

2 The Monocyte - Macrophage Data Set

For our analysis we combined the gene-chip results from three different experimental settings to the monocyte - macrophage (MoMa) data set [7]. In each experiment human peripheral blood monocytes were isolated from healthy donors (experiment 1 and 2) and from donors with Niemann-Pick type C disease (experiment 3). Monocytes were differentiated to macrophages for 4 days in the presence of M-CSF (50 ng/ml,R&D Systems). Differentiation was confirmed by phase contrast microscopy. Gene expression profiles were determined using Affymetrix HG-U133A (experiment 1 and 2) and HG-U133plus2.0 (experiment 3) GeneChips covering 22215 probe sets and about 18400 transcripts (HG-U133A). Probe sets only covered by HG-U133plus2.0 array were excluded from further analysis. In experiment 1 pooled RNA was used for hybridization, while in experiment 2 and 3 RNA from single donors were used. The final data set consisted of seven monocyte and seven macrophage expression profiles and contained 22215 probe sets. After filtering out probe sets which had at least one absent call, 5969 probe sets remained for further analysis. Exp. 1 - 7 refer to monocytes and Exp. 8 - 14 to macrophages. Exp. 1 - 4 and 8 - 11 stem from healthy subjects, the rest from diseased subjects.

3 Methods

3.1 Data Representation and Preprocessing

The gene expression levels are represented by an $(N \times M)$ data matrix $\mathbf{X} = [\mathbf{x}_{*1} \cdots \mathbf{x}_{*M}]$ with the columns $\mathbf{x}_{*m}$ representing the *gene expression signatures* of all genes in one of the M experiments conducted, and the rows represent *gene expression profiles* (GEP) of each of the N genes across all M experimental conditions. Column vectors are denoted as $\mathbf{x}_{*m}$, while row vectors are denoted as

$\mathbf{x}_{n*}$ in the following. The index m is always signifying an environmental condition in the following whereas the index n always refers to a certain gene.

With NMF, a decomposition is sought according to $\mathbf{X} = \mathbf{WH}$. The columns of $\mathbf{W}$ are called *metagenes* and the rows of $\mathbf{H}$ are called *meta-experiments*. Note that the data matrix is non-square with $N \approx 10^3 \cdot M$ which renders a transposition of the data matrix necessary when techniques like ICA are applied. Hence ICA follows the data model: $\mathbf{X}^T = \mathbf{AS}$ where the columns of matrix $\mathbf{A}$ represent the basis vectors of the new representation and may be named *feature profiles*, while the (as much as possible) independent rows of $\mathbf{S}$ are called *expression modes*.

3.2 Feature Selection Schemes

Matrix factorization techniques like ICA or NMF seem promising in generating features suitable for diagnostic classification purposes. In the following we will discuss these techniques in their application to microarray data sets.

ICA - Analysis

The $M \times N$ data matrix $\mathbf{X}^T$ was used as input to the JADE-algorithm [2,3] which encompasses centering and whitening procedures as pre-processing steps. Rather than discussing an ICA-based gene grouping scheme proposed by [5] we propose a different approach which utilizes basic properties of the matrix decomposition model (see Fig. 1) and incorporates diagnostic knowledge to evaluate the structure of the *feature profiles*, the basis vectors of the new representation, and deduce a corresponding *expression mode* to identify marker genes related to the classification task at hand.

Each investigated microarray data set represents at least two different classes of cells, such as cell lines taken either from healthy subjects (class 1) or patients

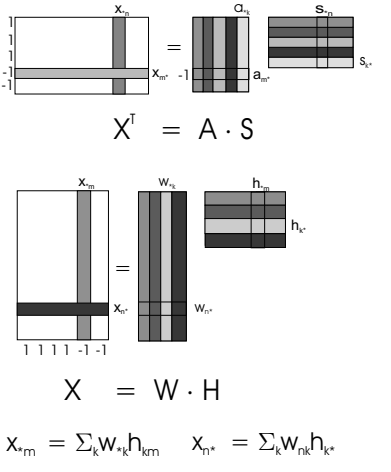


Fig. 1. Illustration of the matrix decomposition scheme used with ICA or NMF

suffering from any disease (class -1). If the *gene expression signatures* $\mathbf{x}_{1*}, \ldots, \mathbf{x}_{M*}$ of the different experiments are arranged according to their diagnostic class label, i.e. the j experiments of class 1 constitute the first j rows of the data matrix $\mathbf{X}^T$, whereas the members of class -1, i.e. $\{\mathbf{x}_{m*}\}_{m=j+1}^{M}$, are collected in the remaining rows, this assignment is also valid for the rows of matrix $\mathbf{A}$. Suppose one of the independent *expression modes* $\mathbf{s}_{m*}$ is characteristic of a putative cellular gene regulation process, which is related to the diagnostic difference between the classes. Then to every *gene expression signature* of class 1, this characteristic *expression mode* should contribute substantially - signalled by a large weight in the corresponding *feature profile* - whereas its contribution to class -1 experiments should be less (or vice versa). Since the k-th column of $\mathbf{A}$ contains the weights with which $\mathbf{s}_{k*}$ contributes to all observed *gene expression signatures*, this column should show large/small feature components according to the class labels. The strategy concentrates on the identification of a column of $\mathbf{A}$, which shows a class specific signature. Informative columns were identified using the correlation coefficient $corr(\mathbf{a}_{*n}, \mathbf{d})$ of each column vector of $\mathbf{A}$ with a design vector $\mathbf{d}$ whose m-th entry is $d_m = \pm 1$, according to the class label of experiment $\mathbf{x}_{m*}$.

Local NMF - Analysis

Applying NMF, the data matrix corresponds to the usual $N \times M$ matrix $\mathbf{X} = [\mathbf{x}_{*1} \ldots \mathbf{x}_{*M}]$. Each column of $\mathbf{X}$, called a gene expression signature, comprises the gene expression levels of all genes resulting from one experiment. After applying the LNMF- algorithm [6], the data matrix is decomposed into two new matrices $\mathbf{W}$ and $\mathbf{H}$. The columns of $\mathbf{W} = [\mathbf{w}_{*1} \ldots \mathbf{w}_{*k}]$ are called *metagenes*. One of them is expected to be characteristic of a regulatory process, which is responsible for the class specific difference in the observed experiments. Its contribution to the observed *gene expression signatures* is contained in the rows of matrix $\mathbf{H}$ which are called *meta-experiments*. Once an informative *meta-experiment* is identified through its correlation to the design vector encompassing the diagnostic information available, further analysis can be focussed only on the genes contained in the corresponding *metagene*. The strategy is similar to the one used in case of the ICA analysis discussed above.

3.3 Matrix Factorization Classifier

Matrix factorization of the data matrix cannot only be considered a feature extraction technique but the component matrices containing either the *expression modes* (ICA: $\mathbf{S}$) or the *meta-experiments* (NMF: $\mathbf{H}$) in its rows can be used directly for classification purposes. This is because the column vectors of these matrices contain the weights with which either the *feature profiles* or the *metagenes* contribute to each observed expression signature (NMF) or expression profile (ICA). The diagnostic information is coded in the labels of the expression profiles forming the rows of $\mathbf{X}$ or the columns of $\mathbf{X}^T$. But these labels transfer to corresponding labels for the columns of $\mathbf{H}$ or the rows of $\mathbf{A}$. Hence

any meaningful decomposition should result in a structure of $\mathbf{A}$ or $\mathbf{H}$ from which the class memberships can be obtained.

The method will be explained in the following referring to the NMF notation but can be translated to the ICA notation immediately by recognizing the correspondence of *metagene* - *expression mode* and *feature profile* - *meta-experiment*, i.e. $\mathbf{W} \triangleq \mathbf{S}^T$ and $\mathbf{H} \triangleq \mathbf{A}^T$, respectively.

Note that from $\mathbf{X} = \mathbf{WH}$ it follows that $\mathbf{W}^{\#} \cdot \mathbf{x}_{*k} = \mathbf{h}_{*k}$ where $\mathbf{W}^{\#}$ denotes the pseudo-inverse of $\mathbf{W}$. Now the similarities between the columns of $\mathbf{H}$ can be used to classify the observations. Though any method based on similarity measures can be used, we simply estimate the correlation coefficient, i.e. $c_m \equiv corr(\mathbf{h}^{test}, \mathbf{h}_{*m})$. Now for each class a separate set S_i of indices is created, where S_1 encompasses all indices m for which $\mathbf{x}_{*m} \in$ class 1 while S_{-1} contains all remaining indices. This thus results in two sets of correlation coefficients corresponding to the two assignments $m \in S_1$ or $m \in S_{-1}$. Two rules for class assignment were tested:

- Average correlation:

$$label(\mathbf{h}^{test}_{*m}) = \begin{cases} 1 \ \langle c_m(1) \rangle > \langle c_m(-1) \rangle \\ -1 \text{ otherwise} \end{cases} \tag{1}$$

  where $\langle .... \rangle$ denotes an average of the correlation coefficients over the respective index set.
- Maximal correlation:

$$label(\mathbf{h}^{test}_{*m}) = \begin{cases} 1 \ max\{c_m(1)\} > max\{c_m(-1)\} \\ -1 \text{ otherwise} \end{cases} \tag{2}$$

  where $max\{c_m(\pm 1)\}$ denotes the maximal value of all correlation coefficients within either the set S_1 or S_{-1}.

Rule 1 thus assigns the class label according to an average correlation of the test vector with all vectors belonging two one or the other index set. Rule 2 assigns the class label according to the maximal correlation occurring between the test vector and the members of each index set.

4 Discussion

4.1 Matrix Decomposition Techniques for Feature Selection

The following results discuss the application of the matrix factorization techniques to identify *expression modes* or *metagenes* relevant to the diagnostic classification. From them we then extract marker genes and identify related regulatory networks. We analyze the set of expression profiles from the monocyte - macrophage differentiation study [7].

ICA Analysis

A systematic investigation of the structure of the mixing matrices was carried out while increasing the extracted number of *expression modes* from $k =$

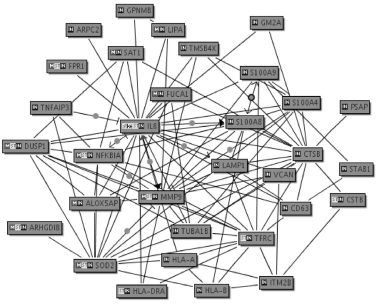


Fig. 2. Gene regulatory network encompassing the marker genes identified with ICA

$2, \ldots, 14$. The resulting maximal correlation coefficients $|corr(\mathbf{a}_{*k}, \mathbf{d}_{*i})|$ showed little variation in both cases with average values $\langle |corr(\mathbf{a}_{*k}, \mathbf{d}_{*1})| \rangle = 0.79$ and $\langle |corr(\mathbf{a}_{*k}, \mathbf{d}_{*2})| \rangle = 0.94$. The maxima occur at $k = 3$ in case 1 and at $k = 8$ in case 2. Considering the 10 most strongly expressed genes in each case, it is found that these marker genes are involved in a gene regulation network exemplified in Fig. 2. For a discussion of the related pathways identified by applying BiblioSphere MeSH- and GeneOntology filter tools see [7].

Table 1. The genes, shown in Fig. 2, could be linked with MeSH term: MHC Class I [G04.610.626.580.595], Z-score(32.6)

NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha
HLA-B	major histocompatibility complex, class I, B
HLA-A	major histocompatibility complex, class I, A
IL8	interleukin 8
MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
TFRC	transferrin receptor (p90, CD71)

The major known pathways associated to M-CSF receptor dependent signalling include expansion of the role of the MAP-kinase pathway and Jun/Fos, Jak/Stat and PI-3 kinase dependent signal transduction. Up-regulation of immune-regulatory components involved in innate immunity response (e.g. MHC), specific and nonspecific opsonin receptors as well as charge and motif pattern recognition receptors is characteristic for monocyte/macrophage differentiation. Beyond this, an increase of membrane biogenesis, vesicular trafficking and metabolic pathways including amino acids, glucose, fatty acids and sterols, as well as increased activity of lysosomal hydrolases that enhance phagocytotic function, autophagy and recycling is triggered through M-CSF signaling as a hallmark of innate immunity. These mechanisms are tightly coupled to changes in cytokine/chemokine response and red/ox signalling that drive chemotaxis migration, inflammation, apoptosis and survival. Tab. 1 and 2 as well as Fig. 2

Table 2. These genes, shown in Fig. 2, could be linked to the following functions

	Signal transduction
CSPG2	chondroitin sulfate proteoglycan 2 (versican)
DUSP1	dual specificity phosphatase 1
S100A8	S100 calcium binding protein A8 (calgranulin A)
S100A9	S100 calcium binding protein A9 (calgranulin B)
TNFAIP3	tumor necrosis factor, alpha-induced protein 3
IL8	interleukin 8
MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
SOD2	superoxide dismutase 2, mitochondrial
	Regulatory sequences
ALOX5AP	arachidonate 5-lipoxygenase-activating protein
GM2A	GM2 ganglioside activator
S100A8	S100 calcium binding protein A8 (calgranulin A)
SOD2	superoxide dismutase 2, mitochondrial
	Survival/Apoptosis
ALOX5AP	arachidonate 5-lipoxygenase-activating protein
CSPG2	chondroitin sulfate proteoglycan 2 (versican)
DUSP1	dual specificity phosphatase 1
LAMP1	lysosomal-associated membrane protein 1
S100A8	S100 calcium binding protein A8 (calgranulin A)
S100A9	S100 calcium binding protein A9 (calgranulin B)
SOD2	superoxide dismutase 2, mitochondrial

summarize the marker genes extracted and show their membership in different pathways and regulatory networks. For a more in depth discussion see [7].

LNMF Analysis

Monocyte vs Macrophage: A LNMF analysis was also performed on the 14 experiments of the MoMa data set. Again the number k of extracted *metagenes* was varied systematically to identify an optimal decomposition of the $N \times M$ data matrix $\mathbf{X}$. For every k, the correlation coefficients between the *meta-experiments* and the design vectors $\mathbf{d}_{i*}, i = 1, 2$ were computed. Figure 3 exhibits the signature of row $\mathbf{h}_{28*}$ of $\mathbf{H}_{k=29}$ and the related *metagene*.

Healthy vs. Diseased: In this case, the number of *meta-experiments* with a strong correlation with the design vector reflecting over-expressed genes in case of cell lines taken from Niemann Pick C patients increases nearly linearly with increasing k. In case of under-expressed *metagenes* related to the disease, only a few significant *meta-experiments* appear for $k > 60$. As an example, a decomposition in $k = 370$ *metagenes* is considered. *Meta-experiment* $\mathbf{h}_{13*}$ yields

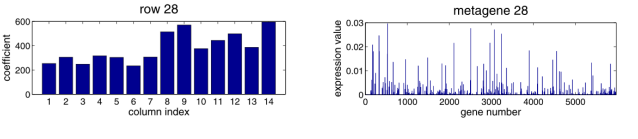


Fig. 3. Signature of row $\mathbf{h}_{28*}$ of $\mathbf{H}_{k=29}$ (*meta-experiment* 28) and corresponding column $\mathbf{w}_{*28}$ of $\mathbf{W}_{k=29}$ (*metagene* 28)

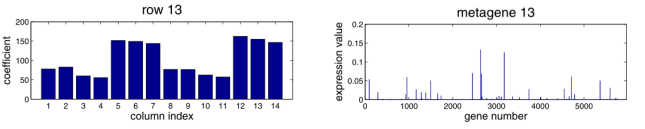


Fig. 4. Signature of *meta-experiment* $\mathbf{h}_{13*}$ and corresponding *metagene* $\mathbf{w}_{*13}$

$corr(\mathbf{h}_{13*}, \mathbf{d}_{2*}) = -0.98$ with respect to the separation between the classes "healthy" and "diseased" (see figure 4).

4.2 Matrix Factorization Classifier

The following results discuss the application of the matrix factorization techniques to extract marker genes and related regulatory networks. classifier to the microarray data sets. The cases to be distinguished by the classifier in the data sets are the following: *case 1*: monocyte vs macrophage and *case 2*: healthy vs Niemann Pick C disease.

MoMa Data Set

Applying the MF-classifier described above, the similarity between either the rows of matrix **A** (ICA) or the columns of matrix **H** (NMF) was studied using the data set with LOO cross-validation. As can be seen from Tab. 3, in most cases one false classification occurred leading to the suspicion of a falsely classified training data sample which could be identified as experiment nr.4. With this

Table 3. Classification errors of the matrix decomposition classifiers using ICA (left) and NMF (right) for matrix decomposition and LOO cross-validation. Class labels: Mo - Monocyte, Ma - Macrophage, he - healthy, NPC - Niemann Pick C.

	JADE		LNMF		JADE		LNMF	
	Mo vs. Ma		Mo vs. Ma		he vs. NPC		he vs. NPC	
k	avg	max	avg	max	avg	max	avg	max
2	11	11	4	9	11	12	3	5
3	3	5	1	1	5	5	0	0
4	3	4	1	1	5	3	0	0
5	2	2	1	1	4	2	0	0
6	3	2	1	1	4	1	0	0
7	1	1	1	1	1	0	0	0
8	1	1	1	1	0	0	0	0
9	1	2	1	1	0	0	0	0
10	2	2	2	1	0	0	0	0
11	1	2	1	1	0	0	0	0
12	1	2	1	2	0	0	0	0
13	1	2	2	2	0	0	0	0
14	1	2	2	2	0	0	0	0

data set, the LNMF algorithm obtained a decomposition of the data matrix which lead to a much more robust classification of all four diagnostic classes underlying the data set with respect to a variation of k.

Comparison of Expression Modes and Metagenes

Though both the ICA and the LNMF algorithm lead to data matrix decompositions which finally lead to robust and efficient diagnostic classification of the data sets, they nonetheless resulted with groups of strongly over- or under-expressed genes in the related metagenes and expression profiles which showed only partial overlap. It is interesting to compare the distribution of correlation coefficients of the individual expression profiles of the identified marker genes for both algorithms. It turned out that the LNMF algorithm results in a much narrower distribution of c-values meaning that it yields a much more consistent set of diagnostic marker genes when measured by the correlation to the diagnostic design vector.

5 Conclusion

In this study an especially promising gene extraction scheme was studied which first transformed the data into a feature space and then concentrated on the most informative feature vector to deduce a group of marker genes to feed into the classifier. Both the feature extraction procedure and the construction of the classifier rely on matrix decomposition techniques and make use of prior diagnostic knowledge to identify the most informative *feature profiles* or *meta-experiments* and their related *expression modes* or *metagenes*. The method performed robustly and efficiently on the data sets studied. The importance of the selected genes could be quantified by measuring the correlation of their *expression profiles* with the diagnostic design vector. Final cross-validation proved the efficiency and reliability of the methods studied in identifying diagnostic marker genes. A closer inspection showed that most extracted marker genes could be identified as members of known pathways of monocyte - macrophage differentiation.

Acknowledgement

Support by grants from Siemens AG, Munich, the DFG (Graduate College 638) and the DAAD (PPP Luso - Alem$\tilde{a}$ and PPP Hispano - Almeanas) is gratefully acknowledged.

References

1. Brunet, J.-P., Tamayo, P., Golub, T., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. PNAS 101, 4164–4169 (2004)
2. Cardoso, J.-F., Souloumiac, A.: Blind beamformimg for non-gaussian signals. IEEE Proc. F140(6), 362–370 (1993)

3. Cardoso, J.-F., Souloumiac, A.: Jacobi angles for simultaneous diagonalization. SIAM Journal Mat. Anal. Appl. 17(1), 161–164 (1996)
4. Kuo, B.C., Chang, K.Y.: Feature extractions for small sample size classification problem. IEEE Trans. Geoscience Remote Sensing 45, 756–764 (2007)
5. Lee, S.-I., Batzoglou, S.: Application of independent component analysis to microarrays. Genome Biology 4, R76.1–R76.21 (2003)
6. Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.: Learning spatially localized, parts-based representation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1 (2001)
7. Lutter, D., Ugocsai, P., Grandl, M., Orso, E., Theis, F.J., Lang, E.W., Schmitz, G.: Analysing M-CSF dependent monocyte/macrophage differentiation and meta-clustering with independent component analysis derived expression modes. BMC Bioinformatics 9, 100 (2008)
8. Schachtner, R., Lutter, D., Theis, F.J., Lang, E.W., Tomé, A.M., Górriz-Saez, J., Puntonet, C.G.: Blind matrix decomposition techniques to identify marker genes from microarrays. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666. Springer, Heidelberg (2007)
9. Spang, R., Zuzan, H., West, M., Nevins, J., Blanchette, C., Marks, J.R.: Prediction and uncertainty in the analysis of gene expression profiles. Silico Biology 2, 33–58 (2002)

Named Entity Recognition and Normalization: A Domain-Specific Language Approach

Miguel Vazquez[1,*], Monica Chagoyen[2,3], and Alberto Pascual-Montano[2]

[1] Departamento de Ingeniería del Software e Inteligencia Artificial
miguel.vazquez@fdi.ucm.es
[2] Dpto. Arquitectura de Computadores, Universidad Complutense de Madrid, Madrid, Spain
[3] Biocomputing Unit, Centro Nacional de Biotecnología - CSIC, Madrid, Spain

Summary. We present, RNer, a tool that performs Named Entity Recognition and Normalization of gene and protein mentions on biomedical text. The tool we present not only offers a complete solution to the problem, but it does so by providing easily configurable framework, that abstracts the algorithmic details from the domain specific. Configuration and tuning for particular tasks is done using domain specific languages, clearer and more succinct, yet equally expressive that general purpose languages. An evaluation of the system is carried using the BioCreative datasets.

1 Introduction

Three factors have motivated the biomedical community to get interested in literature mining [10]. The first one is due to the recent high throughput techniques, like DNA microarrays, that, by involving large collections of entities, place bigger demands in the researcher's use of previous knowledge. The second factor is the fast pace at which scientific articles are being published, making it hard for researchers to keep up to date. Coupled with these is the trend followed by many editorials of making the articles available on-line, thus, making them readily available for analysis by software tools.

Named entity recognition (NER) is an important tool for literature mining, as it is found as an initial step in many information extraction applications [11]. Named entity recognition has been used in domains other than the biomedical, like news wire text. It is in the biomedical domain, however, where it presents the most challenges. Gene and protein names are the amongst the most common subjects for NER in the biomedical domain. They have several characteristics that makes them specially challenging. They are very numerous, in the order of millions, the number growing continuously, and the names used to refer to them are not necessarily standardized [7]. Furthermore, in many cases the names are the same across different genes or match some common English words, producing a significant amount of ambiguity [1]. These problems not only affect finding the mentions in the text, but also the subsequent task of identifying the actual entity

* Corresponding author.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 147–155, 2009.
springerlink.com

they refer to. This identification step is usually referred to as normalization, and is as much a challenging task as NER itself [2].

Several initiatives have dealt with the problem of NER in the biomedical context. We have studied closely the BioCreative competition [11, 2], which had several tasks relating to NER and normalization. We used some of the resources they provided to evaluate our system. As for the availability of tools for NER and normalization, we have found several freely available tools for the NER step alone. Some of them, like Abner and Banner [8, 6] we have studied in detail. At the time we did not find any available tool that implemented the normalization step, which motivated us to develop our own.

Examining the state of the art in NER reveals a strong trend towards the use of conditional random fields (CRF). These probabilistic models have proven to be appropriate for text labeling [5], which is how NER is usually approached. For NER, we use CRF to predict labels for a sequence of words given a set of features associated with each word. Most of the work on NER using CRF differs basically in the features used. In fact, the nature of the CRF algorithm reduces the problem to determining a suitable set of features. Most common features are orthographic: Does the word have any uppercase letter? Any digits? Is there a slash character in the word? Etc.

We developed a system, which we call RNer, for both NER and normalization that implemented a standard approach, but offered a practical way to be extended and configured. For NER, we implemented the same ideas from Abner and Banner, abstracted the core inner workings of the algorithm and committed all the details on creating the features to a domain specific language (DSL). For the normalization task we developed our own solution, based on some ideas from the BioCreative competition, and also separated all the configuration options to domain specific languages. The Ruby programming language was used in the implementations, mainly because its clear syntax and meta-programming possibilities makes defining DSL a very easy task.

The advantage of the use of DSLs is that it allows to integrate and test new ideas, for both NER and Normalization, using a more descriptive language than the original programming language. This makes adding new functionality closer to configuration than it is to programming, which has enabled us to produce a working tool with competitive performance with very little effort.

2 System Overview

This is an overview of the process we have implemented. Following sections will cover each aspect in more detail, but this section should be useful to gain perspective for further discussion.

The CRF model for NER is trained off-line. The training text is turned into a stream of words, with the words forming part of the entity mentions labeled accordingly. The model is trained so that, when a new stream of words comes along, it can assign the labels to determine mention boundaries.

The Normalization system is not trained like the NER, but rather, a data structure is generated from a lexicon file to help identify the mentions. The lexicon file contains each gene, along with the list of identifiers with which it is commonly referred to in the literature, usually in a per-organism basis. The Normalization works by matching the mentions found with the synonyms in the lexicon by progressively transforming each into simpler forms until a match, or matches, are made. The matches so found are then evaluated according to a set of rules to produce a match score. This match score is used to select the best match, or to reject them all, if none is found to be good enough. If two matches have the same score, the system selects the best match by comparing the overlap between the words in the text where the mention was found, and the text describing the gene in the correspondent Entrez GeneRIF entry.

3 Named Entity Recognition

The approach we follow for NER is based on that of Abner. We will describe briefly how it works. A more detailed description can be found in [5] or [9].

Finding mentions in text means determining the words that are likely to constitute a gene name. This is commonly done by labeling the words with their position relative to the mentions. The `IOB` schema that we use [8], has the label `B` mark the initial word on a mention, label `I` mark any other word in the mention, and label `O` is used for words that are outside of any mention. Named Entity Recognition is done by assigning these labels to a new sequence of words.

The labeling of a sequence of words using Conditional Random Fields is done by representing each word by a set of features, and using a probabilistic model to determine the sequence of labels that was most likely to have produce such a sequence of features. This model is built in an off-line, phase using example data, which, in our case, is the one provided as training for the BioCreative competition. Optionally, an specific NER model can be build for each organism by including in the training data the gene synonyms found in the organism lexicon.

Our system uses CRF++ [4] to build the model and produce the labels, as it provides bindings for Ruby. Abner and Banner use Mallet, a Java implementation of CRF with similar characteristics. The main difference between all three systems resides in the features used to represent each word. Abner defines a set of features based on morphological features, these features are further expanded in Banner, which includes semantic features as well. Our system abstracts all the feature generation from the rest of the system, and provides a DSL language to specify it. The features used to evaluate our system include those in Abner and Banner, excluding the syntactic based ones, as they slowed the system down considerably, and they did not seem to significantly enhance performance in our application.

We will now take a look at the DSL used to define the features. Figure 1 shows the definition of three of the features exemplifying the three ways to do this. The first one defines the feature `hasDigits` with a regular expression, so that

```
hasDigits     /\d/
prefix_3      /^(...)/
downcase      do |w| w.downcase end
```

Fig. 1. Feature declaration

if the word has a digit, then the feature is true. In the second case, `prefix_3` is defined with a regular expression capture, in this case the feature value is what ever is captured in by the parenthesis, the first three characters. The last example uses a code block, the result of which, the word in lower case, is assigned to the feature. These three features, along with 25 others, are described in the default configuration file, and can be easily over written, and extended. Regular expressions are used since they themselves are a DSL for string matching, and arguably the best way to capture the requisites of this particular domain.

The CRF tool that we use, CRF++ [4], has a particular characteristic of being able to consider a certain context around each word at any particular step, as opposed to just the current state and word features. This context may be defined in another section of our DSL, as shown in figure 2, which just states that features `special, token2, isPunctuation` and `isDelim` from words surrounding the current one in distance of up three should be considered. The features used here must have been defined in the previous DSL.

```
context_features   %w(special token2 isPunctuation isDelim)
context_window     %w(1 2 3 -1 -2 -3)
```

Fig. 2. CRF++ context

4 Normalization

Normalization is the name given to the process of identifying to what gene or protein a mention found in the text refers to. The BioCreative competition features some methods that do not separate NER from normalization. However the most common choice is to consider both process separate.

Our approach to Normalization is a 3 step pipeline. Mentions are assigned a number of candidate genes. The choice for each candidate is scored, filtered, and sorted; and the best is selected, if any survive the filtering. In the case that a draw still remains after scoring, it is resolved in a so called disambiguation step. Let us look at these three process in more detail: candidate matching, scoring and disambiguation.

Candidate matching is done using a ordered list of string matching functions, each one been more permissive than the previous one. The idea behind this setup is to find the best matches as soon as possible. This early stop avoids generating an unnecessary large list of candidates. However, functions at the end of the list should allow for lax enough matching, so that candidates are produced, even if unlikely; the following scoring step will use a more sophisticated method to

```
equal    do |w| [w] end
standard do |w| [w.downcase.split(/\s+/).sort.join("")] end
words    do |w|  w.scan(/[a-z]+/i)  end
```

Fig. 3. Declaration of name indexes

reevaluate their appropriateness. The matching functions are implemented using cue indexes. Each function turns a gene name into a list of cues, forming an index that associates each cue with the list of genes having the correspondent names. When a mention to a gene is found in the text, the system produces the cues for the first index, and checks if any matches are made; if not, it repeats the process for the next index. This is carried along until a match is made, or the last index also fails.

The system only needs the cue generator functions, which are defined using a DSL like in figure 3. In this abridged example the first index function cue is the name or mention as-is; the second takes each word in the mention, sorts them, and joins them together in lowercase –this is a more accommodating cue; whereas the third index returns a list of cues composed of the words in the name or mention. Having in common any of the cues in the list will be enough to produce a match. For example, the mention `lactamase beta 2` would generate the cues `lactamase beta 2` for the *equal* index, `2betalactamase` for the *standard* index and `lactamase`, `beta`, and `2` for the *words* index.

The candidates generated in the previous process might be numerous, including plenty of false positives. To evaluate how good a match each candidate gene is, we compare each of the synonyms the gene has with the mention, generating similarity scores. The candidate gene is assigned the score of the best ranking name. The score is calculated as follows. Both mention and names are chunked down into tokens, and each token is assigned a token class. Each token class is then examined to evaluate the overlap in tokens. Each overlapping scenario is assigned a positive or negative score, and the total sum is the similarity score. The system uses a DSL to specify the token classes and another DSL to specify the cases and their scores.

The three examples in figure 4 illustrate the three ways to define the token types. The first one identifies a token as a roman numeral using a regular expression. If only the name of the type is given, as done in the second case, a default regular expression is attached, matching the same name, case insensitive, and with a possible "`s`" character at the end, to account for possible plurals (This is offered for convenience, an explicit regular expression could be used for cases that do not adhere to this pluralization rule). The last example uses a code block to make this check, in this particular case, it makes use of a hash of Greek letter

```
roman        /^[IV]+$/
promoter
greek        do |w| $greek[w.downcase] != nil end
```

Fig. 4. Token types

```
same.greek      5
miss.greek      -3
diff.promoter   -10
```

Fig. 5. Token comparison weights

names to check if the token is one of them. This, again, is done case insensitive. Tokens not assigned to any of the defined classes are assigned the class *other*. For example, `Lactamase Beta 2` would have a token of class *number*, which will be 2, a token of class *greek*, which will be `B`, and a token of class *other*, which will be `lactamase`. The tokens proposed in the application by default include 12 rules and 11 additional words.

Example figure 5 shows three ways to specify the scores that receive the different cases that could arise when comparing two token classes. The first states that if the same Greek letter names are found in the two, a value of 5 is added to the similarity measure. The second states that if the mention is missing a Greek letter, 3 is subtracted, and the last one states that if one of them has the token `promoter` and the other does not, 10 is subtracted. We have six operators: `same, distinct, common, diff, miss,` and `extra`. Meaning that they have the same tokens, no tokens in common, at least one in common, at least one different, the mention is missing one or the mention has an extra token. For example, the mention `Lactamase 2` and the name `Lactamase Beta 2` would have a common *other* token, a common *number* token, but the mention would be missing a *greek* token.

```
transform.roman do |t| [t[0].arabic, 'number'] end
compare.number do |l1,l2|
  val = 0
  val -= 4 if (l1 - l2).length >0 || (l2 - l1).length >0
  val -= 8 if l1[0]  != l2[0] &&  l1[0]
  val += 3 if l1[0]  == l2[0]
  val
end
```

Fig. 6. Advanced comparisons: Transformations and custom comparisons

In order to add more flexibility to this method, we have added two other operators: *transform*, used to change tokens from one type to another, and *compare*, that allows comparing a certain type of tokens using a code block. Figure 6 holds an example of both. The first one turns a roman to a number, allowing the number `1` to match `I` in roman form, for example. The second compares the numbers of mention and name in a finner grained manner.

We now have the ability to assign a similarity score between a mention and each of its candidate genes. For each candidate gene the best score amongst those from all it's known synonyms is selected. We use these score to rule out those candidates that score to low, and also to establish a ranking amongst those who score high enough. If there are several candidate genes containing the same

synonym (as it often occurs) they will score the same. In order to resolve the draws, we have a final step of disambiguation.

The disambiguation step is done by comparing the context in which the mention occurs, the paragraph for example, with the description of that gene in the Entrez Gene database. The comparison is made by finding the words in common between the text in the mentions context and the text in the gene description, and adding their word weights. The word weights are their inverse document frequency [3], calculated from a corpus of example text drawn from PubMed article abstracts.

5 Evaluation

The NER system comes with a set of default configurations, mostly copied from Abner, but with a few additions. The system performs fairly well with these defaults for gene mention recognition. We plugged our system into the BioCreative competition evaluation sets using these defaults. The results are shown in table 1, we downloaded Abner and Banner and performed the same tests. RNer, as we will call our system, seems to outperform Abner slightly. Banner performs significantly better than both, possibly because it uses syntactic information as features, which our default configuration did not include.

The Normalization step also comes with a default configuration. We show results for the BioCreative I task 1-B in tables 2 and 3. These results are obtained using only the default configuration, no specific tuning is performed for either organism, nor is the provided training data used in any way. We have performed the analysis using our NER system, as well as Abner and Banner. Our NER

Table 1. Results for the BioCreative I task 1-B

System	Precision	Recall	F-Measure
RNer	0.819	0.780	0.799
Abner	0.789	0.741	0.764
Banner	0.836	0.828	0.832

Table 2. Results for the yeast dataset of the BioCreative I task 1-B

NER	Precision	Recall	F-Measure
RNer	0.936	0.863	0.898
Abner	0.941	0.809	0.870
Banner	0.933	0.816	0.870

Table 3. Results for the mouse dataset of the BioCreative I task 1-B

NER	Precision	Recall	F-Measure
RNer	0.695	0.645	0.669
Abner	0.680	0.649	0.664
Banner	0.666	0.686	0.675

system outperforms Abner and Banner in the yeast dataset. Banner, however, outperforms both in the mouse dataset.

6 Discussion

Ruby, our implementation language, proved to be very useful due to its clear syntax and meta-programming possibilities. Ruby can also be compiled into Java bytecode and used in any Java framework. The result is a simple to use system that works out of the box but at the same time is flexible and easy to configure and adapt to other domains. Domain specific languages allow us to describe the specifics of the system in a more clear and succinct way.

While the system does not, for the time being, beat any record in performance, it shows competitive and promising results, even though no organism based tuning is performed and the training data for normalization, provided in BioCreative, is not being used.

7 Availability

This system was developed as part of another system called SENT, available at `http://sent.dacya.ucm.es`. Very likely, this system will be made available to public domain when SENT is published, along with the rest of the code that composes the SENT. In the meantime it will be provided free upon request.

Acknowledgements

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, CYTED-505PI0058, TIN2005-5619. APM acknowledges the support of the Spanish Ramón y Cajal program.

References

1. Chen, L., Liu, H., Friedman, C.: Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics 21(2), 248–256 (2005)
2. Hirschman, L., Colosimo, M., Morgan, A., Yeh, A.: Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics 6(1), 11 (2005)
3. Jones, K.S., et al.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)
4. Kudo, T.: Crf++: Yet another crf toolkit (2005)
5. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning table of contents, pp. 282–289 (2001)
6. Leaman, R., Gonzalez, G.: Banner: An Executable Survey Of Advance. In: Biomedical Named Entity Recognition. In: Pacific Symposium of Biocomputing (PSB) (2008)

7. Leser, U., Hakenberg, J.: What makes a gene name? Named entity recognition in the biomedical literature. Briefings in Bioinformatics 6(4), 357–369 (2005)
8. Settles, B.: Abner: an open source tool for automatically tagging genes, proteins and other entity names in text (2005)
9. Settles, B., Collier, N., Ruch, P., Nazarenko, A.: Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In: COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004, pp. 107–110 (2004)
10. Shatkay, H., Feldman, R.: Mining the Biomedical Literature in the Genomic Era: An Overview. Journal of Computational Biology 10(6), 821–855 (2003)
11. Yeh, A., Morgan, A., Colosimo, M., Hirschman, L.: BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics 6, 1 (2005)

BIORED - A Genetic Algorithm for Pattern Detection in Biosequences

Pedro Pereira[1], Fernando Silva[1], and Nuno A. Fonseca[2]

[1] CRACS & Department of Computer Science, Faculty of Science, Universidade do Porto
Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal
{pdr,fds}@dcc.fc.up.pt
[2] Instituto de Biologia Molecular e Celular (IBMC) & CRACS, Universidade do Porto
Rua do Campo Alegre 823, 4150-180 Porto, Portugal
nf@ibmc.up.pt

Summary. We present a new, efficient and scalable tool, named BIORED, for pattern discovery in proteomic and genomic sequences. It uses a genetic algorithm to find interesting patterns in the form of regular expressions, and a new efficient pattern matching procedure to count pattern occurrences. We studied the performance, scalability and usefulness of BIORED using several databases of biosequences. The results show that BIORED was successful in finding previously known patterns, thus an excellent indicator for its potential. BIORED is available for download under the GNU Public License at http://www.dcc.fc.up.pt/biored/. An online demo is available at the same address.

Keywords: motif discovery, genetic algorithm.

1 Introduction

The identification of interesting patterns (or subsequences) in biosequences has an important role in computational biology. Databases of genomic and proteomic sequences have grown exponentially, and therefore pattern discovery still is a hard problem requiring clever algorithmic to achieve manageable levels of efficiency and powerful pattern languages to be useful.

Patterns often have an important biological significance, hence pattern discovery is an important problem in computational biology. It is, however, a computationally hard task, given the combinatorial involved. The rationale behind pattern discovery in biosequences (proteomic and genomic) is that the patterns correspond to subsequences preserved through evolution, and the reason for being preserved is because they are important to the function or structure of the molecule.

In this paper we describe BIORED, a new efficient and scalable tool to discover interesting patterns in genomic and proteomic sequences. It accepts a powerful pattern language that is a subset of regular expressions and uses a novel genetic algorithm to discover patterns together with a new efficient pattern matching procedure to count pattern occurrences in the sequences. We validate BIORED by applying it to several databases in order to try to rediscover previous known patterns and we study its sequential performance. BIORED is capable of efficiently finding patterns in very large sequence databases and be used to find considerably large patterns.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 156–165, 2009.
springerlink.com

2 Background

The problem of pattern discovery here addressed can be stated as follows. Let Σ be an alphabet of residues (proteomic or genomic). Given a set of sequences S, each sequence composed with characters not restricted to the alphabet Σ, and a pattern size k, the goal is to find the best interesting pattern p with size k accordingly to some evaluation function.

We consider deterministic patterns with wild-cards and ambiguous characters. More specifically, the pattern language is a subset of regular expressions. Every position in the regular expression can be only composed by classes of characters belonging to Σ. A class is represented within brackets. The "`.`" (referred to as don't care character) is used to denote a class of characters composed by all elements in Σ. For compactness of representation, it is also possible to negate the class. In this case, all characters belonging to the alphabet and not shown in the class, are the ones that compose the class. The negation is denoted by "^". For instance, the pattern with length 3 "`[GT].A`" has two matches in the sequence $AA\underline{TAA}G\underline{TTA}A$.

The chosen pattern language is a compromise between simplicity and power. The idea is to allow the discovery of complex patterns while having a sufficiently fast matching algorithm. Although interesting patterns may have gaps, which may be the result of deletions or insertions, many others have undergone smaller mutations and have an equal length. The principle is that we can usually find sub-patterns of larger patterns and later extend them. Another advantage of using (a subset) of regular expressions is that the resulting language is well supported by a considerable number of programs (e.g., grep, sed, emacs, etc) and programming languages (e.g., Perl, PHP, etc).

3 A Genetic Algorithm for Pattern Discovery in Biosequences

BIORED uses a genetic algorithm (GA) [1] to perform pattern discovery. It receives as input a database containing a set of sequences, the length of a pattern k, and some other parameters (such as the maximum number of generations i), and tries to find an interesting pattern of length k.

The implementation of a GA requires the prior definition of a (1) a genetic representation of a pattern (solution), and (2) a fitness function to evaluate the patterns. The implementation of the fitness function involves counting the number of matches of a pattern in the input sequences. This can be a limiting performance factor for the algorithm, therefore we devised an efficient matching procedure.

We next describe the genetic operators used, the fitness function (interestingness metric) and a sequential algorithm for counting the matches.

3.1 Genetic Operators

A genetic operator [1] is a process that aims at maintaining genetic diversity. The operators are analogous to those that occur in the natural world: survival of the fittest, or selection; sexual or asexual reproduction, or crossover; and mutation.

BIORED implements a rank selection operator that sorts the individuals in the population by comparing their fitness value. Each individual is then given a probability of being chosen for reproduction depending on its position. For n individuals, a $n(n+1)/2$ slots roulette-wheel is constructed, with the fittest individual receiving n slots, the second fittest $n-1$, and so on, with the least fit individual receiving just one slot.

During the alternation (or reproduction) phase of the GA, we use three classical genetic operators: mutation, crossover and elitism. The crossover operator selects a character position in the individual to be generated. It then sets the first part with the contents of the first individual and the second part with the contents of the second individual (both selected using a rank selection operator). The mutation operator randomly flips some of the bits that compose the chromosome. The elitism operator selects some of the best individuals to be copied verbatim to the next generation, without suffering any mutation.

3.2 Interestingness Based on Statistical Significance

To guide the search for a pattern and for ranking a set of patterns one needs some measure to assess, in some way, their quality or interestingness. In a GA context, such measure is called fitness function. In complex problems, such as pattern discovery, GAs have a tendency to converge towards local optima rather than the global optimum of the problem. This problem may be alleviated by using a different fitness function, or by using techniques to maintain a diverse population of solutions. Therefore, two fitness functions were considered based on statistical interestingness.

Several approaches have been proposed to determine if a pattern is statistically interesting [2, 3, 4], i.e., if the number of occurrences of a pattern in a set of sequences is greater than the expected value. A pattern is considered statistically interesting if it is overrepresented in the sequences where it occurs. To measure the over-representation, we need to consider the expected number of occurrences and the standard deviation of this value. Equivalently, we need to know how the values are distributed.

We assumed that the probability of the symbols (from Σ) to appear in S are independent and identically distributed. Under these assumptions, the word probability follows a Binomial distribution. The Binomial distribution gives the discrete probability $b(x;\ n, p)$ of obtaining exactly x successes (matches) out of n Bernoulli trials (pattern positions). We consider every character position, that can be a possible place for the word occurrence, as a Bernoulli trial. For example, if we have the sequence `ACGATCAGTACA` and the pattern that we are computing the statistics for has length 5 then there are exactly 8 places where the pattern can occur. Generalizing, having a sequence and a word of length S_n and W_n respectively, there are $S_n - W_n + 1$ places where the word can appear if $S_n \geq W_n$ or zero otherwise. Each Bernoulli trial is true with probability p. The probability p is the multiplication of the probabilities of the individual pattern positions. In turn, the pattern positions probabilities is the sum of the probabilities of the symbols that compose the position. For efficiency reasons, the binomial distribution is approximated by the Poisson distribution for large values of n and small values of p, with $\lambda = np$, or equivalently $p(x;\ \lambda) \approx b(x;\ n, \lambda/n)$ [5].

We are interested to know if the pattern is overrepresented, therefore we calculate the probability of the pattern to appear at least the same number of times in a database as it

effectively appears. Equivalently, we compute the complementary cumulative distribution function (F_c) of the Poisson distribution for $x-1$: $Z = F_c(x-1) = P(X > x-1)$. Since, the Z can take very small values we use the negative logarithmic of Z, more specifically, $-\log(Z)$. We next denote $-\log(Z)$ as $\mathcal{I}$.

The first fitness function relates the interestingness of the pattern with its complexity,

$$f_1 = \frac{\mathcal{I}}{\text{complexity}^x}$$

for $x = 0, 1, 2, 3$. The *complexity* is the sum of the number of characters recognized by each pattern position. For instance, `ACGT` has complexity 4, while `[AC]CGT` has complexity 5 and `[AC][CG][GT][TA]` has complexity 8. The parameter x is used to reduce the patterns complexity, thus improving their quality. Generally, the low quality patterns are a direct result of being too general.

The second fitness function (f_2) borrows ideas from the evaluation function F-measure,

$$f_2 = \frac{2 \times \text{logpn} \times \text{cpx}}{\text{logpn} + \text{cpx}}$$

where p is the probability of the pattern, m is the maximum complexity with same length and alphabet can have, $\text{cpx} = 1 - \frac{\text{complexity}}{m}$, $\text{logpn} = \mathcal{I}/10000$. A ceiling of 10000 is assumed to the value of $\mathcal{I}$.

In general, it is not possible to determine which fitness function behaves better in a set of sequences without some kind of experimentation. This experimentation needs only to be done once for each sequence, and can be done automatically by executing the programs for all the possible fitness functions and choosing the one that achieves the best results.

3.3 Counting Matches

The fitness function, or interestingness metric, requires knowing the number of (overlapping) occurrences of every pattern in the sequence. For example, in the sequence `AAATAA`, the pattern `AA` occurs three times and the pattern `AA[AT]` occurs twice.

Counting the number of occurrences of a single pattern can be troublesome. For instance, if the sequences have total length of n and the pattern is composed by either symbols or unit-length don't care characters with length m, the best algorithm runs in $O(n \log m)$ time (worst-case) [6]. If we could come up with an algorithm with an equal complexity for the worst-case, the best we could do would be $O(ni \log m)$, where i is the number of different patterns (number of individuals of the population). However, since unit-length don't care characters are a subset of classes of characters, the chosen pattern language is more powerful than the pattern language referred in [6].

Since the GA generates several individuals (patterns) in each generation that need to be evaluated, we tried to devise an efficient method to evaluate them simultaneously.

If the algorithm could only handle a single pattern, then it is possible to use a linear solution based on bit-parallelism [7] if the pattern length is small (only a few machine words are needed). The bits are used to simulate a non-deterministic finite automaton (NFA) that describes the pattern.

To expand the algorithm to evaluate several patterns at once, a window with length k is moved through the sequence. Note that all patterns have the same length k. For each window position every pattern is checked for a match. In a sequence with size n, the number of window positions (window size is k) is $n - k + 1$ (assuming that $n \geq k$).

The counting matches algorithm worst-case complexity is $O(nik)$ with the input size n, i the number of individuals in the population, and k the length of the patterns. However, the algorithm is on average much faster, achieving a complexity of $O(ni/w)$, where w is the number of bits in a machine word. The average complexity is directly linked to the average case of the naive string matching.

In spite of the effort to have an efficient counting operation, it remains the bottleneck of the matching algorithm. A parallel version of BIORED was thus developed that achieved linear speedup up to 22 processors [8].

3.4 Implementation

The BIORED was implemented using the C language because the speed was crucial and to perform an extreme control on memory usage. For the statistic functions we used the R [9] library. Note that BIORED can be executed in a variety of platforms, such as clusters and in GRIDs.

The alphabet letters (representing nucleotides or residues) are implemented using an unsigned integer with 32 bits. This representation has the advantage of being simple to apply the genetic operators, namely the crossover and the mutation. This means that a population with i individuals, each having length k, uses exactly $4ik$ bytes of memory using the DNA alphabet. In general, the algorithm uses $|\Sigma|ik/8$ bytes of memory, where Σ is the alphabet used.

We use a binary vector as a chromosome to represent a pattern. The binary vector can, conceptually, be seen as signaling if a character belonging to Σ is present or not at a determinate pattern position. For example, the DNA pattern `[AC]T[ACGT]G` is represented as `1100,0001,1111,0010`, if `A` is represented with the bit-mask `1000`, `C` `0100`, `G` `0010` and `T` `0001`.

The initial population (set of patterns) is randomly initialized. Each bit in an individual has the probability 0.7 of being activated (this value was selected after performing several experiments). The probability was empirically chosen to guarantee the diversity of the population, representing patterns that actually occur in the data.

The probability of undergoing crossover was set to 0.75 and the mutation probability to 0.01. Only the fittest individual is considered an elite. These values were chosen after some experiments with DNA and residue sequences and are the values that proved to work better. By default, the program halts after completing 500 generations. This value was chosen based on the performance experiments done.

Finally, it is worth mention that BIORED includes an option setting to allow the use of symbol probabilities (distribution) different from those observed in the sequences. This requires the user to give an extra file (with sequences) to the program from where the distribution is computed. An example of the usefulness of this option is demonstrated in Section 5.

4 Performance Evaluation

We study the performance of BIORED and the behavior of the GA in terms of convergence and execution time. The databases used in the experiments are indicated in Table 1 and were obtained from the release 38 of the Ensembl project [10]. All experiments were ran in a Cluster with Dual core "AMD Opteron Processor 250" computers, with 4 gigabytes of RAM (but only 600 Mb free) running the Linux operating system (kernel 2.6).

Table 1. Organisms used for evaluation

Organism	Length (bp)
Saccharomyces cerevisiae (whole genome)	12156606
Anopheles gambiae (chromosome 2R)	61545105
Drosophila melanogaster (whole genome)	144141726

Figure 1 shows the effect on the runtime when we alter a single parameter, such as the population size or the pattern length. Theoretically, the runtime is expected to double when the population size is doubled. However, the optimizations performed in the algorithm makes the runtime vary.

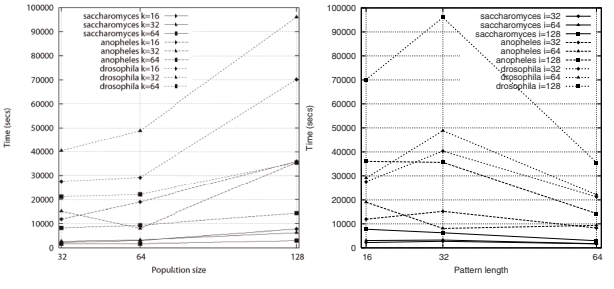


Fig. 1. Run time variation with different populations and pattern lengths (in seconds)

The three organisms used (see Table 1) can be processed in about (largest to smallest) 27, 10 and 4 hours, running for one-thousand generations with a population size of 128 individuals and searching for patterns with length of 64. These values for running times are, in our view, excellent for a sequential execution, considering the relative large size of the data used. In fact, other pattern discovery tools failed to cope with the same data, thus making it impossible to compare relative efficiency (details are discussed in Section 6).

When the pattern length is increased something apparently strange happens. Until a certain pattern length the runtime increases and then it decreases. This is, once again, related to the size of the search space. When the search space grows too much, the genetic algorithm has difficulty in finding an admissible pattern. A possible solution

to this problem could be to initialize the population with statistically interesting words (naturally, found with another tool).

The results show a very small variation on the runtimes when the population sizes increase from 32 to 64 individuals. This effect is a direct consequence of the implemented bit-parallelism technique and the execution on 64-bit architectures.

We evaluated BIORED's convergence and discovered that when the pattern length increases, the population size must also be increased for the convergence to be smoother. This happens because it is more difficult to obtain an admissible large pattern. This was expected, since the search space of a DNA pattern with length 64 is $2^{4\times 64}$. Furthermore, as the size of the pattern decreases, the faster the algorithm converges. This was also expected since the search space becomes exponentially smaller as the size of the pattern decreases.

5 Validation

We demonstrate the usefulness of BIORED in two case studies. The goal is to rediscover some already known patterns.

Human Gene for Proinsulin

In the first case study we chose a database with the human gene for proinsulin from chromosome 11 [11]. BIORED was configured to run with a population of 32 individuals, pattern length of 14, and to stop after one-thousand generations. It yielded the pattern `[CG][AT]GGGG[AT][CG][AT]GGGG[AT]` with a score of 381.6, occurring 48 times and with a probability of 0.00000133. The pattern found is very similar to a previously reported pattern `ACAGGGGTGTGGGG` [12].

Drosophila Melanogaster

In the second case study, we used a database with the whole genome of the Drosophila melanogaster. More concretly, we used the organism disjoint introns (sections of DNA that are spliced out after transcription but before the RNA is used) as input to BIORED, and configured it to use a population of 64 individuals, and a pattern length of 27. The symbols probabilities were gathered from the whole genome. The best pattern after 4096 generations was `ATTGTAAGTCTTTAAATATATTCGTGT` with a score of 7309.4, occurs 256 times and has a probability smaller than 10^{-9}. Curiously, this pattern is a sub-word of the consensus described in [13]. The consensus was manually converted to a regular expression producing (after some simplification) the following pattern: `[^G][AG]AGTT[CT]GT[^A][GT]C[CT]T[AG]AGTCTTT[CT]GTTT`. Note that the original pattern, as it is, achieves a score of 904.2 on the entire genome, i.e., the entire genome was used to compute the distribution of the symbols, while the pattern found by BIORED achieves a score of 7309.4.

6 Related Work

Several pattern discovery tools and algorithms have been developed [14, 15, 4]. Some approaches are based on exhaustive search that guarantee that the best pattern

(accordingly to some specifications) is found. An heuristic approach does not guarantee that the best pattern is found, instead it finds a good "enough" pattern. The advantage of the heuristic approach is that it is often faster than the exhaustive search, but may not find the best solution (pattern).

The Teiresias [14] is closely related to our proposal in terms of the pattern language. It is based on well-organized exhaustive search based on combinations of shorter patterns. The Teiresias algorithm guarantees that all maximal [14] patterns are reported. The algorithm needs to receive as input the minimum number of literals that a pattern can contain, L. Another required parameter is W that indicates the maximum distance between any consecutive L literals. In general, if we use the same L parameter and increase the W parameter, the execution time of the scanning phase, which is the phase where the algorithm gathers seed patterns with the desired L and W characteristics, greatly increases. In the worst case the algorithm is $O(n^3 log\ n)$, but it is reported to work very well when the inputs are highly regular and the parameters W and L are small.

The admissible patterns are similar to the ones we consider. The original Teiresias algorithm only supported one wild card equivalent to our ".". Newer versions support equivalency classes. In an equivalency class the user needs to specify the characters that are to be treated as equal in the actual pattern discovery process. These are similar to the classes of characters supported by BIORED.

A critical problem with Teiresias is that it has a very high memory usage. In an attempt to compare the performance of Teiresias with BIORED we configured Teiresias to support the same classes of characters as BIORED, and tried to identify the previously discovered pattern in the human gene of proinsulin. Teiresias crashed after 8 minutes with a memory consumption of several gigabytes. These parameters were chosen to verify if it could identify the previously discovered pattern in the human gene of proinsulin using the BIORED.

In conclusion, Teiresias seems to be unable to cope with classes that overlap each other (which is exactly the classes supported by BIORED) since it has an extremely high memory consumption that prevents any empirical comparison since it crashes even for small sequences.

Pratt [15] is another tool to discover patterns conserved in sets of unaligned protein sequences. The patterns that can be found are a subset of the patterns that can be described using Prosite notation [16]. In particular, variable length gaps are allowed. Pratt is very memory intensive, contrasting to BIORED, which is pretty light in memory consumption. Pratt tries to find a pattern that occurs in the greatest number of sequences as possible, while the program presented here considers the total number of occurrences in all sequences.

MEME (Multiple EM for Motif Elicitation) [17] uses a stochastic search to discover patterns. It does not require a pattern length parameter, which can be estimated by the algorithm itself. The algorithm is based on expectation maximization technique. Individual MEME patterns cannot contain gaps, and thus are equivalent to the patterns we consider. The overall complexity of MEME is quadratic in the size of the database and linear in the length of the pattern [17], while our proposal is linear in the size of the database and in the length of the pattern.

7 Conclusion

We presented a new pattern discovery tool that discovers interesting patterns, in the form of a regular expression. using a genetic algorithm. The algorithm has a conservative memory usage of $O(ik|\Sigma|)$ and a worst-case time complexity of $O(nikg)$, where Σ is the alphabet used, i is the number of individuals of the population, k is the length of the pattern, n is the size of the input, and g is the number of generations. However, the algorithm is on average much faster, achieving a complexity of $O(gni/w)$, where w is the number of bits in a machine word. The average complexity is directly linked to the average case of the naive string matching. Experiments showed the usefulness of the algorithm, by demonstrating that it is capable of discovering previously known patterns.

The contributions of this paper are three-fold. First, we describe and evaluate a tool that uses a genetic algorithm to discover patterns in genomic and proteomic sequences. Second, we also propose an efficient pattern matching procedure, a crucial component for achieving high performance in any pattern discovery tool. Finally, for a practitioner we provide a pattern discovery tool that is efficient (in terms of execution time and memory usage), has a powerful pattern language, does not impose restrictions on the pattern length, is general (can handle proteomic and genomic sequences), and can handle large databases of sequences, and can be used in a with number of settings (personal computers, clusters, and grids).

Finally, there is still space for improvement. For instance, BIORED implements a simple and fast statistical approach to determine the interestigness of a pattern. In order to improve the statistical accuracy, we plan to include, as an option, more rigorous tests such us the recently proposed complementary statistical tests for accessing exceptionalities of motif counts [18].

Acknowledgements

This work has been partially supported by projects ILP-Web-Service (PTDC / EIA / 70841 / 2006), JEDI (PTDC / EIA / 66924 / 2006), STAMPA (PTDC / EIA / 67738 / 2006) and by Fundação para a Ciência e Tecnologia. Nuno Fonseca is funded by the FCT grant SFRH/BPD/26737/2006 and Pedro Pereira by the FCT grant SFRH/BD/30628/2006.

References

1. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs (Third, Revised and Extended edn.). Springer, New York (1999)
2. van Helden, J., del Olmo, M., Perez-Ortin, J.: Statistical analisys of yeast genome downstream sequences reveals putative polyadenylation signals. Nucleic Acids Research 28(4), 1000–1010 (2000)
3. Sinha, S., Tompa, M.: A statistical method for finding transcription factor binding sites. Proceedings of the National Academy of Sciences of the United States of America 95(6), 2738–2743 (2000)

4. Sinha, S., Tompa, M.: An exact method for finding shor motifs in sequences, with application to the ribosome binding site problem. In: Proceedings of the 7th International Conference on ISMB, pp. 262–271 (1999)
5. Feller, W.: An Introduction to Probability Theory and Its Applications, 3rd edn. John Wiley & Sons, Chichester (1968)
6. Cole, R., Hariharan, R.: Verifying candidate matches in sparse and wildcard matching. In: STOC 2002: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, pp. 592–601. ACM Press, New York (2002)
7. Navarro, G.: Pattern matching. Journal of Applied Statistics 31(8), 925–949 (2004); Special issue on Pattern Discovery
8. Pereira, P., Fonseca, N.A., Silva, F.: A high performance distributed tool for mining patterns in biological sequences. Technical Report DCC-2006-08, DCC-FC & LIACC, Universidade do Porto (2006)
9. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2005) ISBN 3-900051-07-0
10. Hubbard, T., Andrews, D., Caccamo, M., et al.: Ensembl 2005. Nucleic Acids Research 33(1) (January 2005)
11. Rotwein, P., Yokoyama, S., Didier, D.K., Chirgwin, J.M.: Genetic analysis of the hypervariable region flanking the human insulin gene. The American Journal of Human Genetics (1986)
12. Lew, A., Rutter, W.J., Kennedy, G.C.: Unusual dna structure of the diabetes susceptibility locus iddm2 and its effect on transcription by the insulin promoter factor pur-1/maz. Proceedings of the National Academy of Sciences of the United States of America 97(23), 12508–12512 (2000)
13. Costas, J., Vieira, C.P., Casares, F., Vieira, J.: Genomic characterization of a repetitive motif strongly associated with developmental genes in drosophila. BMC Genomics (2003)
14. Rigoutsos, I., Floratos, A.: Combinatorial pattern discovery in biological sequences: The teiresias algorithm. Bioinformatics 14(1), 55–67 (1998)
15. Jonassen, I., Collins, J.F., Higgins, D.: Finding flexible patterns in unaligned protein sequences. Protein Science 4(8), 1587–1595 (1995)
16. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M., Sigrist, C.J.A.: The prosite database. Nucleic Acids Res., 34 (2006)
17. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on ISMB, pp. 28–36. AAAI Press, Menlo Park (1994)
18. Robin, S., Schbath, S., Vandewalle, V.: Statistical tests to compare motif count exceptionalities. BMC Bioinformatics 8(84) (2007)

A Recursive Genetic Algorithm to Automatically Select Genes for Cancer Classification

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Safaai Deris[2], and Michifumi Yoshioka[1]

[1] Department of Computer Science and Intelligent Systems,
Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
mohd.saberi@sig.cs.osakafu-u.ac.jp, omatu@cs.osakafu-u.ac.jp,
yoshioka@cs.osakafu-u.ac.jp
[2] Department of Software Engineering,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia
safaai@utm.my

Abstract. Gene expression technology namely microarray, offers the ability to measure the expression levels of thousands of genes simultaneously in a biological organism. Microarray data are expected to be of significant help in the development of efficient cancer diagnosis and classification platform. The main problem that needs to be addressed is the selection of a small subset of genes that contributes to a disease from the thousands of genes measured on microarray that are inherently noisy. Most approaches from previous works have selected the numbers of genes manually and thus, have caused difficulty, especially for beginner biologists. Hence, this paper aims to automatically select a small subset of informative genes that is most relevant for the cancer classification. In order to achieve this aim, a recursive genetic algorithm has been proposed. Experimental results show that the gene subset is small in size and yield better classification accuracy as compared with other previous works as well as four methods experimented in this work. A list of informative genes in the best subsets is also presented for biological usage.

Keywords: Cancer Classification, Gene Selection, Genetic Algorithm, Microarray Data, Recursive Genetic Algorithm, Support Vector Machine.

1 Introduction

The problem of cancer in this world is a growing one. A traditional cancer diagnosis relies on a complex and inexact combination of clinical and histopathological data. This classic approach may fail when dealing with atypical tumours or morphologically indistinguishable tumour subtypes. Advances in the area of microarray-based gene expression analysis have led to a promising future of cancer diagnosis using new molecular-based approaches. This microarray technology allows scientists to measure the expression levels of thousands of genes simultaneously in a biological organism. It produces microarray data as the final product. By using the technology, a comparison between the gene expression levels of cancerous and normal tissues can be done. This comparison is useful to select those genes that might anticipate the clinical behaviour of cancers. Thus, there is a need to select informative genes that contribute to

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 166–174, 2009.
springerlink.com

a cancerous state. An informative gene is a gene that is useful for cancer classification. However, the gene selection for the cancer classification poses a major challenge because of the following characteristics:

- The data sizes are large, for example, the number of genes, *M* and the number of samples, *N* are in the ranges of 10,000-20,000 and 30-200, respectively.
- Most genes are not relevant for classifying different tissue types.
- These data have a noisy nature.

To overcome the problems, a gene selection approach is usually used to select a subset of informative genes for cancer classification [1]. It has several advantages:

- It can maintain or improve classification accuracy.
- It can yield a small subset of genes and reduce dimensionality of the data.
- It can reduce the cost in a clinical setting.

Gene selection methods can be classified into two categories [1]. If a gene selection method is carried out independently from a classification procedure, it belongs to the filter approach. Otherwise, it is said to follow a hybrid (wrapper) approach. In the early era of microarray analysis, most previous works have used the filter approach to select genes since it is computationally more efficient than the hybrid approach [2],[3],[4],[5]. They evaluate a gene based on its discriminative power for the target classes without considering its correlations with other genes. This mechanism may result in inclusion of irrelevant genes in a selected gene set. A few years ago and recently, several hybrid approaches especially a combination between genetic algorithms (GAs) and a classifier, have been implemented to select informative genes [6],[7],[8],[9],[10]. The hybrid approach usually provides greater accuracy than the filter approach since the genes are selected by considering and optimising correlations among genes.

A hybrid of a GA and a support vector machine classifier (GASVM), and an improved GASVM (GASVM-II) have been proposed by our previous work for selecting informative genes [10]. However, these hybrid methods have several limitations. This paper proposes a recursive algorithm based on GASVM-II. This algorithm is called as a recursive genetic algorithm (R-GA). It is developed to improve the performances of GASVM and GASVM-II. The diagnostic goal is to develop a medical procedure based on the least number of possible genes that needed to detect diseases. Thus, the ultimate aim of this paper is to select a small subset of informative genes from microarray data by using R-GA in order to produce higher cancer classification accuracy.

2 GASVM-II

GASVM-II (NewGASVM) has been developed by our previous work [10] to improve the performance of GASVM. Two main components of GASVM-II are GA and support vector machine (SVM) classifiers. Generally, the GA selects subsets of genes and then the SVM classifier evaluates the subsets during a classification process. The overall procedure is shown in Fig. 1.

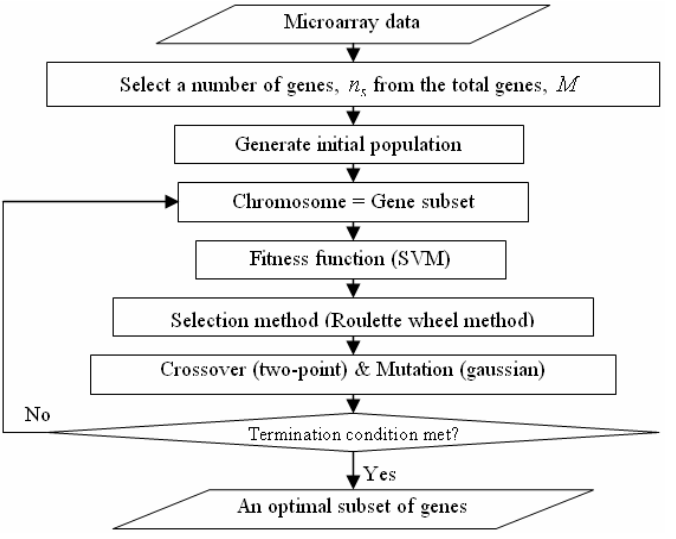


Fig. 1. The flow chart of GASVM-II

g_1	g_2	g_3	...	g_{n_s-2}	g_{n_s-1}	g_{n_s}

Note:
M = the total number of genes in a data set, n_S = a number of selected genes, $1 \le n_S \le M$,
j = the jth gene in a chromosome, $1 \le j \le n_S$, g_j = an integer value in a chromosome, $1 \le g_j \le M$.

Fig. 2. Chromosome representation in GASVM-II

An individual or a chromosome represents a gene subset. In GASVM-II, the chromosome representation is modified to support higher-dimensional data (microarray data) as shown in Fig. 2 which has integer representation. It includes values of an integer, g_j that indicate which genes are needed to be selected among the total number of genes. For example, if $g_j = 10$, then GASVM-II selects the 10th gene from the data set, and groups it into a related subset of genes. The number of selected genes is represented by n_s. The number of g_j is equal to n_s that is fixed at the early stage of the hybrid method. Different from conventional methods, the chromosome is not encoded with all the genes, but with only a fixed number of genes to construct an individual or a solution (a gene subset). Hence, the total of genes, M does not really affect the size (length) of the chromosome so as to keep its size relatively small. Its length can vary according to M and n_s. Finally, the chromosome is represented as $x = (g_1, g_2, ..., g_{n_s})$. For example, the ath chromosome is represented by $x_a = (g_{a,1}, g_{a,2}, ..., g_{a,n_s})$.

A fitness value of an individual (a gene subset) is calculated based on multi-objective optimisation (MOO) that uses two criteria: the accuracy on the training data and the number of selected genes. The theory and application of the MOO for gene

selection and cancer classification can be found in Mohamad *et al*. [6]. The fitness function of an individual is formulated as follow:

$$fitness(x) = w_1 \times A(x) + (w_2(M - R(x)) / M) \quad (1)$$

in which $A(x) \in [0,1]$ is the leave-one-out-cross-validation (LOOCV) accuracy on the training data using the only expression values of selected genes in a subset, x. This accuracy is provided by an SVM classifier. $R(x)$ is the number of selected genes in x. M is the total number of genes. w_1 and w_2 denote two weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$. GASVM-II is used to maximise the fitness function in order to find an optimal gene subset, which has achieved the highest LOOCV accuracy with the smallest number of selected genes.

3 The Proposed Recursive Genetic Algorithm (R-GA)

GASVM and GASVM-II have several advantages and disadvantages [10]. The advantage of GASVM is that it can automatically select a number of genes to produce an optimal gene subset. However, it performs poorly in higher-dimensional data (microarray data) due to its binary chromosome representation. GASVM-II was designed to surmount the limitation. GASVM-II performs well in the higher-dimensional data since it can reduce the complexity of search space and may be able to evaluate all possible subsets of genes. Nevertheless, it selects a number of genes manually and yields inconsistent results when it is run independently. It would be difficult and time consuming to process since microarray data have thousands of genes.

This paper proposes R-GA based on GASVM-II to automatically select a small subset of genes for cancers classification. It has components such as chromosome representation, fitness function, selection method, and operators as implemented in GASVM-II. SVM is used as a classifier since GASVM and GASVM-II have also applied SVM, so that objective comparison between them can honestly be done. Basically, this algorithm repeats the process of GASVM-II to reduce dimensionality of data recursively. In each cycle, a number of genes are automatically reduced and an optimal subset is produced. Furthermore, the complexity of the search or solution space can also be decreased on a cycle by cycle basis. It is repeated until the number of selected genes in the optimal subset is equal to 1. Next, the best subset among the optimal subsets is selected based on the highest fitness value (the highest LOOCV accuracy with the smallest number of selected genes). Finally, the best subset is used to construct an SVM classifier. The constructed SVM is then tested by using the test set. Further details of the proposed R-GA are shown in Fig. 3. Ambroise and Mclachlan have indicated that testing results could be overoptimistic, caused by the "selection bias" if the test samples were not excluded from the classifier building process in a hybrid approach [11]. Therefore, the proposed R-GA has totally excluded the test samples from the classifier building process in order to avoid the influence of bias.

```
VARIABLE / INPUT:
   gen _ num : number of generation.      c : cycle. x_a : ath chromosome.
   # gene : number of genes               S_c : an optimal subset of genes of cycle c.
   div _ gene : divider for the number of selected genes.      gen : generation.
   pop _ num : number of population.      S_best : the best subset of genes.
Begin
   gen := 0;              c := 0;              n_s := M / div _ gene;
   S_c := 0;              S_best := 0;
  while ( S_best.# gene > 1) do
        for (a = 1; a ≤ pop _ num; a + +)
              x_a := initialise(x_a, n_s);
        end_for
        while (gen < gen_num) do
            for (a = 1; a ≤ pop _ num; a + +)
                x_a.fitness := w_1 × A(x_a) + (w_2(M − R(x_a)) / M);
                if (x_a.fitness > S_c.fitness) then
                    S_c := x_a;
                end_if
            end_for
            selection _ method(roulette _ wheel);
            crossover(two _ point);
            mutation(gaussian);          gen := gen + 1;
        end_while
        return (S_C);
        if (S_c.fitness > S_best.fitness) then
              S_best := S_c;
        end_if
        if (S_c.#gene > 100) then
           n_s := S_c.#gene / div _ gene;
           if (n_s < 100) then
              n_s := 100;
           end_if
        end_if
        else if (10 < S_c.#gene ≤ 100) then
              n_s := S_c.#gene − 10;
        end_else_if
        else if (1 < S_c.#gene ≤ 10) then
              n_s := S_c.#gene − 1;
        end_else_if
        S_c.fitness:=0;          c := c + 1;        gen := 0;
  end_while
  return (S_best);
End
```

Fig. 3. The pseudo-code of R-GA

4 Experiments

4.1 Data Sets

Two microarray data sets are used to evaluate R-GA: Lung cancer and mixed-lineage leukaemia (MLL) cancer. The lung cancer data set has two classes: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). There are 181 samples, originally analysed by Armstrong *et al.* [2]. The training set contains 32 of them. The rest of the 149 samples are used for the test set. Each sample is described by 12,533 genes. The MLL cancer data set is a multi-classes data set. It has three leukaemia classes: acute lymphoblastic leukaemia (ALL), acute myeloid leukaemia (AML), and MLL. The training set contains 57 samples, whereas the test set contains 15 samples. There are 12,582 genes in each sample [3].

4.2 Experimental Setup

Three criteria following its importance are considered to evaluate the performances of R-GA and others experimental methods: test accuracy, LOOCV accuracy, and the number of selected genes. #Selected Genes and Run# represent the number of selected genes and a run number, respectively.

Experimental results presented in this paper pursue four objectives. The first objective is to show that gene selection using R-GA is needed to produce a small subset of selected genes for better classification accuracy. The second objective is to display a list of informative genes in the best subsets founded by R-GA for biological usage. The third objective is then to show that R-GA is better than the others experimental methods (GASVM-II, MOGASVM, GASVM, and SVM classifiers), whereas the last objective is to highlight that it is superior to previous works. To achieve the four objectives, several experiments are conducted 10 times on each data set using R-GA. Next, an average result of the 10 runs is obtained.

4.3 Experimental Result

Table 1 shows that the highest accuracies of LOOCV and test for classifying the lung cancer samples are 100% and 98.66.12%, respectively, while 100% LOOCV accuracy and 100% test accuracy are recorded for the MLL data set. The averages of LOOCV accuracy, test accuracy, and the number of selected genes for the lung data were 100%, 93.69%, and 2.8 genes, respectively, whereas 100% LOOCV accuracy, 91.33% test accuracy, and 12.0 selected genes of the MLL data set.

Only four genes in the best subset were finally selected to obtain the highest accuracy of the lung data set, whereas 20 genes were of the MLL data set. These subsets were being chosen as the best subsets for both data since they had the highest test accuracy with the smallest number of selected genes. Interestingly, all runs have achieved 100% LOOCV accuracy. This has proven that R-GA has searched and selected the optimal solution (the best gene subset) successfully in the solution space. However, results of LOOCV accuracy in both the data sets were much higher than test accuracy due to the over-fitting of the data. For the overall results, a gene selection process using the proposed R-GA is needed to produce a small gene subset and obtain better cancers classification.

Table 1. Classification accuracies for each run using R-GA

Run#	Lung Data Set			MLL Data Set		
	LOOCV (%)	Test (%)	#Selected Genes	LOOCV (%)	Test (%)	#Selected Genes
1	100	94.63	2	100	100	20
2	100	93.96	5	100	93.33	8
3	100	94.63	2	100	93.33	20
4	100	90.60	2	100	86.67	20
5	100	93.96	5	100	80.00	10
6	100	98.66	4	100	86.67	10
7	100	94.63	2	100	93.33	8
8	100	94.63	2	100	93.33	8
9	100	90.60	2	100	93.33	8
10	100	90.60	2	100	93.33	8
Average ± S.D	100 ± 0	93.69 ± 2.52	2.80 ± 1.32	100 ± 0	91.33 ± 5.49	12.0 ± 5.58

Note: Results of the best subsets shown in shaded cells. S.D. denotes the standard deviation.

The informative genes in the best gene subsets as founded by R-GA are listed in Table 2. Some of these genes are already identified to be highly possible clinical markers for cancer diagnosis by biological researches. Evans *et al*. reported that the gene U03877 was differentially expressed 2-fold or more in adenoma subtypes [12]. Next, the gene U70063 was identified by Li *et al*. as the top 23 highly significant molecular signatures for sub-typing acute leukaemia [13]. Guinn *et al*. have listed the gene U87459 as an antigen gene [14]. The patent of the United States entitled "Methods and Compositions for the Identification, Assessment, and Therapy of Human Cancers" (Patent number: 7338758; Publication date: March 4, 2008) has found the gene D50930 as highly expressed and sensitive genes. Furthermore, the gene Y00638 was identified by Armstrong *et al*. and their patent in the United States (Patent number: 7011947; Publication date: March 14, 2006) as an over-expressed gene in MLL compared to ALL. Some of the remaining genes may be the excellent candidates for further clinical investigation.

According to Table 3, R-GA has outperformed the other experimental methods in terms of LOOCV accuracy, test accuracy, and the number of selected genes.

Table 2. The list of informative genes in the best gene subsets

Data Set	Run#	Probe-set Name	Gene Accession Number	Gene Description
Lung	6	32551_at	U03877	EGF-containing fibulin-like extracellular matrix protein 1
		33634_at	AF038007	ATPase, Class I, type 8B, member 1
		35708_at	W27414	Homo sapiens, clone IMAGE:3502329, mRNA, partial cds
		36938_at	U70063	N-acylsphingosine amidohydrolase (acid ceramidase)
MLL	1	33636_at	U87459	Human autoimmunogenic cancer
		34129_at	AB023223	Homo sapiens mRNA for KIAA1006 protein, partial cds
		34583_at	U02687	Human growth factor receptor tyrosine kinase (STK-1) mRNA, complete cds
		34441_at	AF052090	Homo sapiens clone 23950 mRNA sequence
		37832_at	AL080062	Homo sapiens mRNA; cDNA DKFZp564I122 (from clone DKFZp564I122)
		38601_at	AF073500	Homo sapiens bestrophin (VMD2) mRNA, alternatively spliced product, complete cds
		39212_at	AF038179	Homo sapiens clone 23939 mRNA sequence
		35659_at	U00672	Human interleukin-10 receptor mRNA, complete cds
		40143_at	D50930	Human mRNA for KIAA0140 gene, complete cds
		40520_g_at	Y00638	Human mRNA for leukocyte common antigen (T200)
		41750_at	D49489	Human mRNA for protein disulfide isomerase-related protein P5, complete cds
		33863_at	U65785	Human 150 kDa oxygen-regulated protein ORP150 mRNA, complete cds
		35804_at	AB022785	Homo sapiens ASH2L gene, complete cds, similar to Drosophila ash2 gene
		41594_at	M64174	Human protein-tyrosine kinase (JAK1) mRNA, complete cds
		326_i_at	HG1800	Ribosomal Protein S20

MOGASVM and GASVM cannot produce an optimised subset of informative genes because they perform poorly in higher-dimensional data due to its chromosome representation limitation. SVM classifier cannot classify cancer samples accurately since it is applied without any gene selection methods. Furthermore, GASVM-II selects the number of genes manually. This manual selection is not practical to be used in real application, and causes difficulty in the usage. On the contrary, R-GA selects the number of genes automatically, removes irrelevant genes, and performs well to obtain the optimised subset.

In Table 4, based on LOOCV and test accuracies of the lung data set, it was noted that the best results (100% LOOCV accuracy and 98.66% test accuracy) from this work were equal to the best result from the latest previous work [7]. However, this work only uses four genes to achieve the accuracies, whereas eight genes have been used in the work of Shah and Kusiak [7]. The first original work [3], achieved only 97.32% test accuracy by using four genes. Similarly, there were increases in LOOCV accuracy (100%) and test accuracy (100%) for the MLL data set as compared to the other previous works [2,4,5]. Moreover, the number of selected genes was also lower than the previous works. Overall, this work has also outperformed the related previous works on both the data sets.

Table 3. Benchmark of the proposed R-GA with the other experimental methods

Method	Lung Data Set (Average ± S.D; The Best)			MLL Data Set (Average ± S.D; The Best)		
	#Selected Genes	Accuracy (%)		#Selected Genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
R-GA	(2.80 ± 1.32; 4)	(100 ± 0; 100)	(93.69 ± 2.52; 98.66)	(12.0 ± 5.58; 20)	(100 ± 0; 100)	(91.33 ± 5.49; 100)
GASVM-II	(10 ± 0; 10)	(100 ± 0; 100)	(59.33 ± 29.32; 97.32)	(30 ± 0; 30)	(100 ± 0; 100)	(84.67 ± 6.33; 93.33)
MOGASVM	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)	(4,465.2 ± 18.34; 4,437)	(94.74 ± 0; 94.74)	(90 ± 3.51; 93.33)
GASVM	(6,267.8 ± 56.34; 6,342)	(75 ± 0; 75)	(84.77 ± 2.53; 87.92)	(6,298.8 ± 51.51; 6,224)	(94.74 ± 0; 94.74)	(87.33 ± 2.11; 86.67)
SVM	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)	(12,582 ± 0; 12,582)	(92.98 ± 0; 92.98)	(86.67 ± 0; 86.67)

Note: The best result shown in shaded cells. S.D. denotes the standard deviation.

Table 4. Benchmark of the R-GA with related previous works

Author [Reference]	Lung Data Set			MLL Data Set		
	#Selected Genes	Accuracy (%)		#Selected Genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
Our work	4	100	98.66	20	100	100
Shah and Kusiak [7]	8	100	98.66	-	-	-
Gordon *et al.* [3]	4	-	97.32	-	-	-
Li *et al.* [4]	-	-	97.99	-	-	100
Wang *et al.* [5]	-	-	-	39	100	-
Armstrong *et al.* [2]	-	-	-	100	95	-

Note: The best result shown in shaded cells. '–' means that result is not reported in the related previous work.

5 Conclusion

In this paper, R-GA has been developed and analysed for gene selection of two microarray data sets. Based on the experimental results, the performance of R-GA was

superior to the other experimental methods and related previous works. This is due to the fact that it can automatically reduce dimensionality of the data. The complexity of search or solution spaces can also be automatically decreased on a cycle by cycle basis. When the dimensionality or complexity was reduced, the combination of genes in the data was also decreased. Thus, the gene selection using the proposed algorithm is needed to produce a small subset for better cancer classification of microarray data. Even though the proposed algorithm has classified tumours with higher accuracy, it is still not able to completely avoid the over-fitting problem. A combination between filter approaches and a hybrid approach is recently developed to solve the problem.

References

1. Mohamad, M.S., Omatu, S., Deris, S., et al.: A model for gene selection and classification of gene expression data. J. Artif. Life Robotics 11(2), 219–222 (2007)
2. Armstrong, S.A., Staunton, J.E., Silverman, L.B., et al.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genetics (30), 41–47 (2002)
3. Gordon, G.J., Jensen, R.V., Hsiao, L.L., et al.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. 62, 4963–4967 (2002)
4. Li, J., Liu, H., Ng, S.K., et al.: Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics 19, 93–102 (2003)
5. Wang, Y., Makedon, F.S., Ford, J.C., et al.: HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21(8), 1530–1537 (2005)
6. Mohamad, M.S., Omatu, S., Deris, S., et al.: A multi-objective strategy in genetic algorithm for gene selection of gene expression data. J. Artif. Life Robotics (appear in press)
7. Shah, S., Kusiak, A.: Cancer gene search with data-mining and genetic algorithms. Computers Bio. Med. 37(2), 251–261 (2007)
8. Huang, H.L., Chang, F.L.: ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. Biosystems 90, 516–528 (2007)
9. Moon, H., Ahn, H., Kodell, R.L., et al.: Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artif. Intell. Med. 41, 197–207 (2007)
10. Mohamad, M.S., Deris, S., Illias, R.M.: A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray. J. Comput. Intell. Appl. 5, 91–107 (2005)
11. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. In: National Academy of Science of the USA, Washington, pp. 6562–6566 (2002)
12. Evans, C., Young, A.N., Brown, M.R., et al.: Novel patterns of gene expression in pituitary adenomas identified by complementary deoxyribonucleic acid microarrays and quantitative reverse transcription-polymerase chain reaction. J. Clin. Endocrinol Metab. 86(7), 3097–3107 (2001)
13. Li, X., Rao, S., Wang, Y., et al.: Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Res. 32(9), 2685–2694 (2004)
14. Guinn, B., Gilkes, A.F., Woodward, E., et al.: Microarray analysis of tumour antigen expression in presentation acute myeloid leukaemia. Biochem. Biophys. Res. Communications 333(3), 703–713 (2005)

On Mining Protein Unfolding Simulation Data with Inductive Logic Programming

Rui Camacho[1], Alexssander Alves[1], Cândida G. Silva[2], and Rui M.M. Brito[2]

[1] LIAAD & Faculdade de Engenharia, Universidade do Porto, Portugal
{rcamacho,aalves}@fe.up.pt
[2] Chemistry Department, Faculty of Science and Technology
and Center for Neuroscience and Cell Biology
University of Coimbra, Portugal
csilva@student.uc.pt, brito@ci.uc.pt

Summary. The detailed study of folding and unfolding events in proteins is becoming central to develop rational therapeutic strategies against maladies such as Alzheimer and Parkinson disease. A promising approach to study the unfolding processes of proteins is through computer simulations. However, these computer simulations generate huge amounts of data that require computational methods for their analysis.

In this paper we report on the use of Inductive Logic Programming (ILP) techniques to analyse the trajectories of protein unfolding simulations. The paper describes ongoing work on one of several problems of interest in the protein unfolding setting. The problem we address here is that of explaining what makes secondary structure elements to break down during the unfolding process. We tackle such problem collecting examples of contexts where secondary structures break and (automatically) constructing rules that may be used to suggest the explanations.

Keywords: Inductive Logic Programming, Protein Unfolding.

1 Introduction

In recent years, the identification of many human and animal diseases as protein misfolding disorders highlighted the importance of the protein folding problem, i.e. the process of conversion of a linear sequence of amino-acids into a functional tri-dimensional structure of a protein. After decades of efforts, this still is an unsolved problem in structural molecular biology. Central in health matters, it is today believed that protein unfolding events are responsible for triggering amyloidogenic processes in several proteins. These processes are at the origin of such disorders as Alzheimer, Parkinson, Bovine Spongiform Encephalopathy (BSE), Familial Amyloid Polyneuropathy (FAP) and several other acquired and hereditary diseases. Thus, the detailed study of folding and unfolding events in proteins is not only important to the characterization of the mechanisms associated with several amyloid diseases but also is becoming central to the development of rational therapeutic strategies against these diseases. In this context, computer

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 175–179, 2009.
springerlink.com

simulations based on molecular dynamics have been succesfully applied to explore and analyse the folding and unfolding events in proteins [5, 4, 1].

The opportunities that datamining methodologies offer to analyse, compare and contrast multiple protein unfolding simulations from different structural classes of proteins, and from amyloidogenic and non-amyloidogenic protein variants, opens new possibilities to allow the production of new knowledge or new views on the protein folding problem and its relationship with health and disease. Finding biologically significant rules may have important repercussions related to human and animal health, because a better understanding of the properties that make a protein amyloidogenic might help in the fight against this debilitating family of diseases - the amyloid diseases.

In order to find differences in the unfolding pathways of amyloidogenic (Am) and non-amyloidogenic(non-Am) variants of transthyretin (TTR), a protein associated with FAP, we use Inductive Logic Porgramming (ILP), a Multi-Relational Data Mining algorithm.

The main advantages of using ILP over competing technologies are sustained by the powerful expressive language to describe both data and the models. This powerful expressiveness as two major consequences: complex models may be constructed to "explain the data"; and the models are generally comprehensible, thus contributing to an insight on the phenomena that produced the data. Furthermore, ILP systems allow domain experts to provide almost any kind of information (ex., structured information like graphs) that may be helpful for the construction of the models. ILP systems may also combine in the same model symbolic relations with numerical computations.

The rest of the paper is organised as follows. Section 2 gives a brief introduction to Inductive Logic Programming. In Section 3 we describe the ongoing experimental work and preliminary results. Section 4 presents the conclusions on the preliminary work and points out future work.

2 ILP in a Nutshell

Inductive Logic Programming (ILP) is a major field in Machine Learning with important applications in Multi-Relational Data Mining. The fundamental goal of a predictive ILP system is to construct models (usually called hypotheses) given background knowledge and observations (usually called examples in the ILP literature).

The task aim is to induce a logic program that given a set of positive and negative examples of the concept to learn, and some prior knowledge (or *background knowledge*), entails all positive examples and no negative example. In the context of this paper, positive examples correspond to events where protein secondary structure break, while the negative examples correspond to instants where there is no break on the secondary structure.

See [3] for an in-depth overview on ILP and [2] for a list of applications.

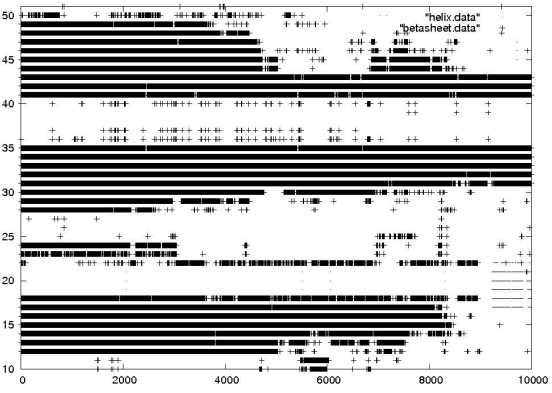


Fig. 1. Evolution of the secondary structure of WT-TTR during one unfolding simulation. The xx axis represents simulation time. The yy axis represents the position of a residue in the protein (only positions between 10 and 51 are represented). Thick lines indicate that the residue belongs to a beta sheet, thin lines indicate that the residue belong to a alpha-helix. We can see that (on top of the picture) the beta sheet between position 41 and 50 looses a substantial amount of residues near simulation time 5000. We can also see that a alpha-helix appears near simulation time 9000 between positions 17 and 23 (near the bottom right side of the picture).

3 Preliminary Experiments

The experiments we have designed have two objectives. First, to find rules that predict the circumstances in which secondary structures break down. Second, find rules that differentiate the break down process in the WT (wild type) and L55P protein variants. This later goal is very important since it may contribute to an explanation for the malignant behaviour of L55P-TTR.

We have used data from simulations of the protein transthyretin (TTR)[1]. The simulations [1] include 5 runs using the wilde type (WT) and 5 others using the amyloidogenic type (L55P). In each simulation run we collected, for each residue and instant of time, information concerning the secondary structure it belongs to and its Solvent Accessible Surface Area (SASA) value. Each SASA file has almost 7.3 MB and each secondary structure information has nearly 2.3MB. The total amount of data produced by the 10 simulations is nearly 100 MB. We have used the Aleph [6] ILP system.

From the original simulation data we constructed the ILP data set as follows. For each secondary structure we take its composition at instant 0 as a reference. Then we trace the simulation looking for an instant where a percentage (system parameter) of residues are no longer part of the structure. That instant marks a positive example. We also store *a window*[2] of simulation traces immediately

[1] Reference 1TTA in the PDB (http://www.rcsb.org/pdb/home/home.do)
[2] The size of the window is also a parameter for the filtering procedure.

```
[Rule 1]  ssBreak(A,B,C,D,E) :-
              sasaSum(A,C,D,E,F), lteqSasa(F,40.0).
(``A structure breaks if the sum of SASA of their residues is < 40'')

[Rule 2]  ssBreak(A,B,C,D,E) :-
              sasaSumVariation(A,1,C,D,E,F), lteqDeltaSasa(F,-50).
(``A structure breaks if the sum of SASA of their residues decreases more than
   or equal to 50 from one instant to the next in the simulation'')

[Rule 3]  ssBreak(A,B,C,D,E) :-
              secStructure(D,sheet,C,F,G), gteqSize(G,10),
              sasaMinValue(A,C,D,E,H), gteqSasa(H,0.1).
(``A beta sheet breaks if its size is greater than 10 residues and all its
   residues have a SAS greater than or equal to 0.1'')
```

Fig. 2. Rules found by Aleph to predict the breakdown of secondary structure of proteins WT-TTR and L55P

preceding this event. This information is stored as ILP background knowledge and may be useful to explain the breaking of secondary structure. The simulation trajectories where there is no secondary structure break are also stored and a sample[3] of them is collected to construct the negative examples. With this filtering procedure we construct the positive and negative example's file and part of the background file (the one containing simulation information).

Apart from information concerning the simulation trajectories, the background knowledge includes a set of predicates useful to construct the models. So far we have encoded and used three major groups of predicates: predicates on the SASA value of residues; predicates on variation of SASA values and; general purpose relational predicates. In the first group we have predicates that compute the sum, the average, maximum value and minimum value of SASA of the residues in the structure. Predicates of the second group compute variations of the previous measures. The third group has predicates to compare numerical quantities.

So far, we have found a small set of rules from which the most accurate are shown in Figure 2. These rules have good individual accuracy and are very easy to interpret. However they only cover ("explain") 53% of the positive examples – events where a secondary structure break. This value suggests that we need to improve the background knowledge, that is, we need more background predicates describing features necessary to "explain" the process.

4 Conclusions and Future Work

In this paper we have described an ILP-based approach to the automatic analysis of protein unfolding simulation data. We have addresses the specific problem of predicting the context where a protein secondary structure will break. Predictive rules were induced by the ILP system Aleph. The rules constructed so far are very easy to understand by the domain experts. On the other hand we have not yet been able to construct a set of rules that explain all the events where

[3] Another system parameter.

secondary structures break. This latter result suggest that further extensions to the background knowledge are required. We have also not yet found interesting rules that discriminate between the wilde type and the amyloidogenic variant of the protein.

Acknowledgments

This work has been partially supported by projects "Searching for high level rules in protein folding and unfolding: from amyloid diseases to protein structure prediction" (PTDC/BIA-PRO/72838/2006) and "ILP-Web-Service" (PTDC/EIA/70841/2006) and the doctoral fellowship SFRH/BD/16888/2004 (to CGS) by Fundação para a Ciência e Tecnologia.

References

1. Brito, R.M.M., Dubitzky, W., Rodrigues, J.R.: Protein folding and unfolding simulations: A new challenge for data mining. OMICS: A Journal of Integrative Biology 8, 153–166 (2004)
2. Dzeroski, S.: Relational Data Mining. Springer, New York (2001)
3. Muggleton, S., De Raedt, L.: Inductive logic programming: Theory and methods. Journal of Logic Programming 19/20, 629–679 (1994)
4. Pande, V.S., Baker, I., Chapman, J., et al.: Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers 68, 91–109 (2003)
5. Shea, J.E., Brooks, C.L.: From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. Annu. Rev. Phys. Chem. 52, 499–535 (2001)
6. Srinivasan, A.: The Aleph Manual (2003), http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph

A Knowledge Discovery Method for the Characterization of Protein Unfolding Processes

Elisabeth Fernandes[1], Alípio M. Jorge[1,2], Cândida G. Silva[3], and Rui M.M. Brito[3]

[1] FEP, University of Porto
[2] LIAAD/ INESC Porto L.A., Portugal
[3] Chemistry Department, Faculty of Sciences and Technology, and Centre for Neuroscience and Cell Biology, University of Coimbra, Portugal
elisabeth.ferna@gmail.com, amjorge@fep.up.pt, csilva@student.uc.pt, brito@ci.uc.pt

Abstract. This work presents a method of knowledge discovery in data obtained from Molecular Dynamics Protein Unfolding Simulations. The data under study was obtained from simulations of the unfolding process of the protein Transthyretin (TTR), responsible for amyloid diseases such as Familial Amyloid Polyneuropathy (FAP). Protein unfolding and misfolding are at the source of many amyloidogenic diseases. Thus, the molecular characterization of protein unfolding processes through experimental and simulation methods may be essential in the development of effective treatments. Here, we analyzed the distance variation of each of the 127 amino acids C_α (alpha carbon) atoms of TTR to the centre of mass of the protein, along 10 different unfolding simulations - five simulations of WT-TTR and five simulations of L55P-TTR, a highly amyloidogenic TTR variant. Using data mining techniques, and considering all the information of the 10 runs, we identified several clusters of amino acids. For each cluster we selected the representative element and identified events which were used as features. With Association Rules we found patterns that characterize the type of TTR variant under study. These results may help discriminate between amyloidogenic and non-amyloidogenic behaviour among different TTR variants and contribute to the understanding of the molecular mechanisms of FAP.

Keywords: Unfolding Protein, Diseases, Data Mining, Knowledge Discovery.

1 Introduction

Understanding the protein folding process is a problem that plays a pivotal role in the post genomic era of molecular biology [17] because it remains impossible to reliably predict the 3-D structure of a protein from its linear sequence of amino acids. Moreover, errors in the folding process are at the source of many diseases, and among them are amyloid disorders like Alzhaimer's and Parkinson's diseases.

In this work we study the unfolding process of the protein *Transthyretin* (TTR), responsible for such pathologies as *Familial Amyloid Polyneuropathy* (FAP), *Familial Amyloid Cardiomyopathy* (FAC) and *Senil System Amyloidosys* (SSA) [6]. The data was generated from molecular dynamics unfolding simulations of TTR, in order to characterize the unfolding process of different variants of TTR and to identify amino

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 180–188, 2009.
springerlink.com

acid residues with coordinated behavior. In this study, we have focused on a particular molecular property, the distance between each one of the 127 amino acid C_α (alpha carbon) atoms and the centre of mass of TTR, along 10 different unfolding simulations. Each simulated process runs for 10 *nanoseconds* (*ns*) of simulation time and protein structures are saved every *picosecond (ps)* corresponding to 10,000 time steps (frames). Since the protein has 127 amino acids and we are measuring one numeric property, each run generates 127 time series of length 10,000. The data set analyzed contains 10 runs, five of which are of WT-TTR (the wild type variant, which tends to unfold correctly), and the other five of L55P-TTR, a highly amyloidogenic TTR variant, where a proline is replacing a leucine in position 55).

Such a complex data set (10,000 × 127 × 10 data points), requires sophisticated methods of analysis to exploit more data while reducing information loss. In this work we present a method of knowledge discovery, involving different techniques, that allows us to extract useful knowledge from such large data sets [8]. The method is based on three steps: the first one is data reduction, where we identify clusters of amino acids and calculate the representative ones; the second one is data modeling, where we move from low-level time series to high-level events; finally, we relate events with the type of the proteins using association rules.

In a previous work, using data from a single simulation process, Azevedo *et al.* [3] used association rules to identify clusters of hydrophobic amino acids. The property under study then was SASA (solvent accessible surface area). More recently, Ferreira *et al.* [9] studied the same SASA property using hierarchical clustering.

2 Molecular Dynamics (MD) Protein Unfolding Simulations

Detailed simulations of the unfolding process allow us to understand and study this phenomenon. In this work we use data generated by simulation of the folding process of TTR and we use classical mechanisms to calculate forces and velocities. The classical MD is based on Newton's equations of motion. Through the integration of these equations we obtain a trajectory that describes positions, and velocities of atoms in the system and how they vary with time [18].

2.1 Simulation Details

Initial coordinates for TTR were obtained from the crystal structure (PDB entry 1tta [13]) and hydrogen atoms were added. All minimization and MD procedures were performed with the program NAMD [15], using version 27 of the CHARMM force field [16]. All atoms were explicitly represented. The programs Dowser [18,20] and Solvate [11] were used to introduce and control water molecules and Na^+Cl^- ions around the protein. The complete system was comprised of 45.000 atoms. The system was minimized, equilibrated and headed to the target temperature under Langevin dynamics.

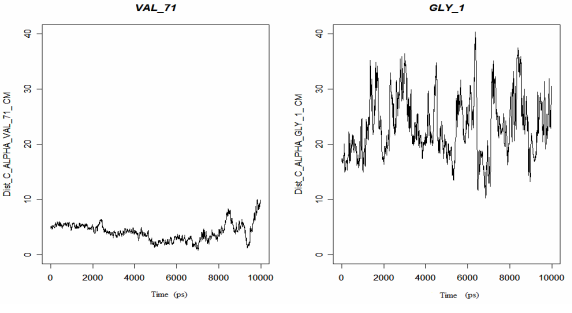
A control simulation was made at 310 K, and several unfolding simulations at 500 K were produced during 10 *nanoseconds* (*ns*). The simulations were carried out using periodic boundary conditions and a time step of 2 *fentoseconds* (*fs*), with distances between hydrogen and heavy atoms constrained. Short range non-bonded interactions were calculated with a 12 Å cut-off, and along range electrostatic interactions were treated using the particle mesh Ewald summation (PME) algorithm [1,6].

2.2 The Data

To analyze the behavior of the amino acids along multiple unfolding simulations of TTR we calculated the values of the distance between the alpha-carbon ($C_{\square}$) of each amino acid and the centre of mass of TTR. We call this property *dtc* (distance to the centre). The monomer of the TTR has 127 amino acids. Each trajectory is a time series with 10,000 frames (each frame was recorded every *picosecond* (*ps*)).

To characterize the unfolding routes we analyze 5 runs of the WT-TTR (wild-type TTR) and 5 runs of the L55P-TTR, a highly amyloidogenic TTR variant. In study we have a three-dimensional data set of 10 simulations, each one with 127 time series (1 per amino acid) along 10,000 frames. Each value in the data set records the *dtc* in run *r*, at time *t* for the aminoacid *a*.

The values of the recorded distance range from 0.14 Å to 50.33Å. Figure 1 shows two examples of observed time series, for the first unfolding simulation of WT-TTR: VAL_71 is always near the centre of TTR (Fig. 1(a)), and GLY_1 presents values of distance above 10 Å (Fig. 1(b)).



(a) (b)

Fig. 1. Variation of *dtc*, the distance between the centre of mass of TTR and the alpha–carbon (C_{α}), of the amino acids VAL_71 (a) and GLY_1 (b), along one unfolding of WT-TTR

3 The Discovery Method

In this section we present the Knowledge Discovery Process we have developed. The first step consists on Clustering and Data Reduction. We started by applying a hierarchical agglomerative clustering over the 10 × 127 time series of *dtc* variation to identify groups of amino-acids that have similar behavior. For each cluster obtained we determined a representative amino acid. This way, the number of amino acids, in study, was reduced from 127 to 15.

After that we looked for interesting events involving the selected amino acids. Using the events as features that characterize each of the 10 runs, we searched for association rules that represent relationships between sets of events and the kind of TTR's variant (see Fig. 2).

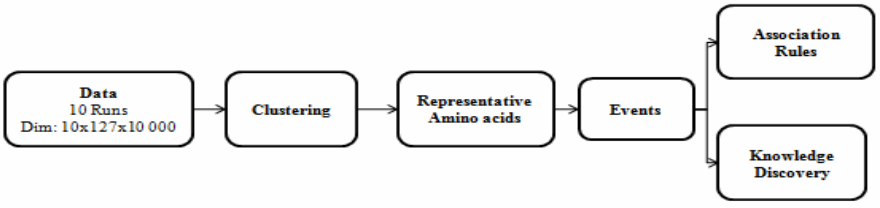


Fig. 2. KDD Method to analyze the molecular mechanisms of TTR. In this work we explore association rules, but other knowledge discovery algorithms can be employed.

3.1 Clustering and Data Reduction

In this first step, the aim is to obtain one partition that groups the amino acids with similar behavior along the simulation. This will allow us, in a subsequent step, to reduce the number of amino acids to focus on. Since 5 of the 10 simulations have a different amino acid on position 55 with respect to the other 5, the time series for this position were excluded. For each simulation we applied hierarchical agglomerative clustering with the Ward´s method (which, at each stage, joins the pair of clusters which leads to minimum increase in the total within-cluster sum of squared distances) and the similarity measure was a plain Euclidean distance. Each amino acid is described by 10,000 features corresponding to the time steps (frames).

To learn the best number of clusters for each partition we calculated and optimized the indexes of Milligan and Cooper (Calinski and Harabasz's, C-Index)[9]. All these procedures were computed using R statistical package (*stat* package [21] and *fpc* package [14]). Given t hierarchical clusterings of the same set of n objects, we can obtain a single *consensus tree* [9]. A consensus tree synthesizes the information contained in the 10 clusterings and gives a summary of the relationships between the amino acids. For that, we have used hierarchical clustering again, but defining a new distance function for the aminoacids, combining the results of the basic clusterings. Thus, for determining the consensus tree, the similarity measure between two amino acids A and B, is the number of times that they stay in the same cluster in each of the 10 initial clusterings. This measure ranges between 0 and 10. We have used average linkage.

As a result, we have obtained 14 clusters of amino acids (Fig. 3). We note that this final clustering takes into account all the 10 simulations. After grouping the amino acids by similarity we identify the representative ones. Our choice was: for each run, the candidate representative amino acid of each cluster is the one closer to the cluster centroid. The final representative element chosen is the most voted amino acid for that cluster. For example, if the amino acid A_i, in ten runs, is more often a representative than the others of the same cluster, we choose A_i to represent that cluster.

Tie cases are left to be solved in the end. In those cases we choose the element that is not close to any representative amino acids selected before on the polypeptide chain. The goal is to obtain a group of representative amino acids that are uniformly distributed on the polypeptide chain, in order to represent the behavior of all amino acids of the chain.

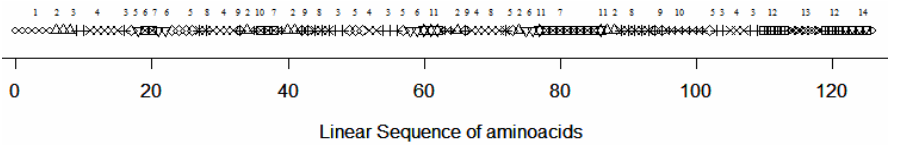


Fig. 3. Representation of the clusters obtained in the Linear Sequence of the 127 amino acids of TTR. Each dot is one amino acid. Each symbol represents one of the 14 clusters**.** Clusters are labeled by numbers above the sequences of dots (amino acids) from the same clusters. For example, cluster 2 contains subsequences of aminoacids around positions 8, 67, 76 and 90.

The amino acids chosen to be the representative were: GLY_4, MET_13, LEU_17, PRO_24, ARG_34, GLU_42, GLU_63, SER_77, LYS_80, ALA_91, ASN_98, SER_112, SER_117, ASN_124. The amino acid at position 55 (which was excluded of the clustering analysis), was also selected as a representative, as it seems to play a key role in the process.

3.2 Event Detection

One of the main goals of this work is to find singular characteristics and interesting phenomenon on temporal series of the amino acids in study. In this section, we describe how we looked for relevant events along the 10 runs. The time series of each amino acid were analyzed, *i.e.*, we searched for significant changes of their behavior along the unfolding simulations. A significant change is called an event, "*a dynamic phenomenon whose behavior changes enough over time to be considered a qualitatively significant change*" [12].

Events Involving Two Amino Acids

In our work we have considered more than one type of event. In this paper we describe results with one of them only, which consists on changes of position between two amino acids with respect to the centre of TTR. To identify these events, we consider the difference vector *df* for each pair of representative amino acids. An event occurs when there is a sequence of at least *n* positive (negative) values in *df* and a posterior sequence of at least *n* negative (positive) values. (see Fig. 4 and Fig. 5)

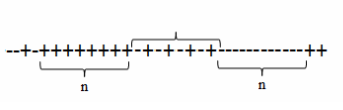


Fig. 4. An event occurs when there is a sequence of *n* equal signs in the difference vector of two amino acids followed (not necessarily immediately) by a sequence of *n* opposite signs

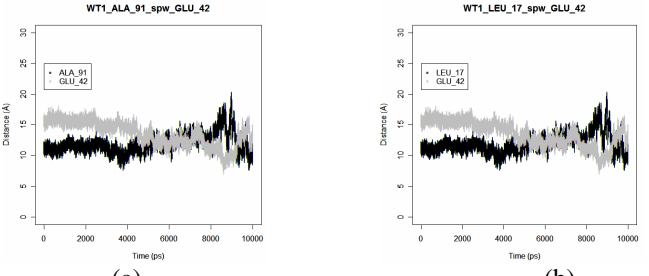
In Fig. we can observe two events of this type, which were found in the first run of WT-TTR. During the first 2000 *ps*, the amino acid ALA_91 was farther away of the centre of TTR than GLU_42, and after 8000 *ps* they change position in relation to the centre of TTR (*switches places with - spw*). The same thing happens with LEU_17

and GLU_42. In short, we move from low-level time series to high-level events so that we can perform data mining on the high level events.

We found 95 events. The most frequent ones are shown on Table 1: **SER_117 spw GLU_42** and **SER_117 spw SER_112** events occur in all WT-TTR runs. **ASN_98 spw SER_77** and **ASN_124 spw GLU_63** occur in all L55P-TTR.

Table 1. Most frequent events

Events	WT_Freq	L55P__Freq	Total _Freq
SER_117 spw GLU_42	**5**	3	8
SER_117 spw SER_112	**5**	3	8
ASN_124 spw LYS_80	4	4	8
ASN_124 spw SER_117	4	4	8
ASN_124 spw GLU_63	2	**5**	7
ASN_98 spw SER_77	2	**5**	7



(a) (b)

Fig. 5. Two occurrences of the specified event. (a) When the distance between LEU_17 at the centre of TTR decrease, the distance between GLU_42 and the centre of TTR increase, the same phenomenon happens with LEU_17 and GLU_42 in the first simulation of WT-TTR.

3.3 Association Rules

Association rules [1] are commonly defined as follows: Let $I = \{i_1,\ldots, i_m\}$ be a set of distinct items, and D a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$. (X is the antecedent and Y is the consequent of the rule).

The quality and strength of an association rule $X \rightarrow Y$ are quantified mainly by the measures *support*, *confidence* and *lift* [5]. Support is the proportion of transactions in D that contain X and Y. It measures the statistical significance of the pattern. Confidence is the fraction of transactions containing X that also contain Y. Lift measures the interestingness of the rule as the quotient between its confidence and the support of its consequent Y. In other words, lift tells us how many times the probability of observing Y is multiplied if we also observe X.

To obtain rules we have used CAREN, a java based implementation of APRIORI that finds frequent patterns based on depth first expansion with bitwise representation [2]. In our case, each of the 10 runs corresponds to one transaction *T*. The items in each *T* are the events that occurred in the respective run, as well as the tag that identifies the run as L55P or WT.

We considered rules with *support* higher than 0.4 and *confidence* higher than 0.9. The only rule obtained was: **ASN_98_spw_SER_77 & ASN_124_spw_GLU_63 → L55P,** with *support* 0.5, *confidence* 1, *lift* 2.This rule indicates that, in the available simulations, the probability of being in presence of the variant **L55P-TTR** is 100%, if we observe that:

- the amino acid ASN_98 is more distant from the TTR centre than SER_77 during 300 frames and then change their relative position;
- and the amino acid ASN_124 is more distant from the TTR centre than GLU_63 during 300 frames and then change their relative position;

The lift value is 2, which means that the probability of having a L55P-TTR variant is twice as much if we observe the events ASN_98_spw_SER_77 & ASN_124_ spw_GLU_63.

We have, therefore, identified two events in the folding process that, together, occur solely in the case of the L55 variant. It is also true that these two events never occur together in the unfolding of the WT variant. These discriminant features have been uncovered only after the reduction of the original data set.

Given that our rule was obtained from 10 runs only, one important question arises: *what is the probability of obtaining an equally good rule in a random way?* This probability can be calculated using the hypergeometric distribution [19]:

$$m(R) = \sum_{i=p}^{\min(t,P)} P(\text{of t cases selected at random, exactly i are in class c}) = \sum_{i=5}^{\min(5,5)} \frac{\binom{5}{i}\binom{5}{5-i}}{\binom{10}{5}} = 0.0198$$

This probability is relatively low. This means that if we randomly generate events, and repeat the discovery method 50 times, we would obtain approximately once a similarly good rule. This is not an ideal situation, but we could apply the same discovery process to a much larger number of runs, if available. However, the amount of available data is limited by the computational cost of the simulation.

4 Conclusions and Future Work

In this work we present a knowledge discovery process that allowed us to extract useful information from 10 molecular dynamics unfolding simulations of the protein Transthyretin. Using the data generated from the simulations, we could obtain a partition of 127 amino acids of TTR, i.e., we condensed all the information contained in 10 simulations and we found a consensus tree that summaryzes the relationships between amino acids. With the partition in 14 groups we calculated the cluster's representatives, and we reduced the complexity of the problem, as we started with 127 amino acids and finished with 15 (14 representatives plus the amino acid at position 55).

By considering distances of pairs of amino acids to the centree of mass of the protein we found 95 events of one given type. This way, we move from low-level time series to high-level events. Association rule mining allowed us to obtain one rule that relates the identified events with the variant TTR.

In the future, this type of rules may become useful in discriminating amyloidogenic and non-amyloidogenic behavior in protein unfolding simulations of several proteins. We believe that the presented discovery method has some characteristics that make it truly interesting. It may be easily applied to a much larger number of simulations (it is highly scalable), and each step of this process can be executed with different options: Other clustering algorithms may be used and the validation of clusters can be made using other indexes; the representative amino acids can be obtained using different criteria. Additional types of events may be defined and looked for. Moreover the time of occurrence of the events may also be taken into account.

Acknowledgements

The authors acknowledge to the project PTDC/BIA-PRO/72838/2006, FCT and FEDER, Portugal and the Centre for Computational Physics, Departamento de Física, Universidade de Coimbra, for the computer resources provided for the MD simulations. We also thank Maria Paula Brito and Paulo Azevedo for technical comments on the work.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large data bases. In: ACM SIGMOD Int'l Conf.On Management of Data, Washington, USA. ACM Press, New York (1993)
2. Azevedo, P.J.: CAREN A java based Apriori implementation for classification purposes. Technical Report, Universidade do Minho: Departamento de Informática (2005)
3. Azevedo, P., Silva, C., Rodrigues, J., Ferreira, N., Brito, R.: Detection of Hydrophobic Clusters in Molecular Dynamics Protein Simulations Using Association Rules. In: Oliveira, J.L., Maojo, V., Martín-Sánchez, F., Pereira, A.S. (eds.) ISBMDA 2005. LNCS (LNBI), vol. 3745, pp. 329–337. Springer, Heidelberg (2005)
4. Berry, M.J.A., Linoff, G.S.: Mastering Data Minig (2000)
5. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. In: ACM SIGMOD/PODS 1997, pp. 265–276 (1997)
6. Brito, R., Dubitzky, W., Rodrigues, R.: Protein Folding and Unfolding Simulations A New Challenge for Data Mining. A Journal of Integrative Biology 8(2), 153–166 (2004)
7. Fayyad (2), U., Piatetsky-Shapiro, Padhraic, S.: From Data Mining to Knowledge Discovery in Databases. In: Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996)
8. Fayyad, U., Piatetsky-Shapiro, G.: The KDD Process for Extracting Useful knowledge from Volumes of Data. Communications of the ACM 39(11), 27–34 (1996)
9. Ferreira, P.G., Silva, C., Brito, R., Azevedo, P.J.: A Closer Look on Protein Unfolding Simulations through Hierarchical Clustering. In: Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology - CIBCB, Hawai, USA, pp. 461–468 (2007)

10. Gordon, A.D.: Classification, 2nd edn. Chapman & Hall /CRC, Boca Raton (1999)
11. Grubmiller, Helmut.: Solvate 1.0. (1996) (accessed, 2007), http://www.mpibpc.mog.de/groups/grubmueller/start/software/solvet/docu.html
12. Guralnik, V., Srivastava, J.: Event Detection from Series Data. In: KDD 1999. Department of Computer Science, University of Minnesota, San Diego (1999)
13. Hamilton, J., Steinrauf, A., Braden, L.K., Liepnieks, B.C., Benson, J., Holmgren, M.D., Sandgren, G., Steen, O.,, L.: The X-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val-30-Met variant to 1.7 A resolution. J. Biol. Chem. 268, 2416–2424 (2003)
14. Hennig, C.: Package fpc Version 1.1-1 (accessed on 2007), http://cran.rproject.org/wen/packages/fpc/index.html
15. Kalé, L., Skeel, R., BBhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Philips, J., Shinozaki, A., Varadarajan, K., Schulten, K.: NAMD2: Greater Scability for Parallel Molecular Dynamics. Journal of Computational Physics 151, 283–312 (1999)
16. MacKerell, A.D., Bashford, D., Bellot, M., Dunbrack, R.L., Evanseck, J.: All-atom empirical potencial for molecular modeling and dynamics studies of proteins. J.Phys. Chem. B 102, 3586–3616 (1998)
17. Pande, V.S., Baker, I., Chapman, J., Elmer, S.P., Khaliq, S., Larson, S.M., Rhee, Y.M., Shirts, M.R., Snow, C.D., Sorin, E.J., Zagrovic, B.: Atomistic protein Folding Simulations on the Submillisecond Time Scale Using Worldwide Distributed Computing. Biopolymers 68, 91–109 (2003)
18. Scheraga, H., Khalili, M., Liwo, A.: Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. Annu. Rev. Phys. Chem., 57–83 (2007)
19. Witten, I., Frank, E.: Data Mining: practical machine learning tools abd techniques with Java implementatons, p. 177. Morgan Kaufman Publishers, San Francisco (1999)
20. Zhang, L., Hermans, J.: Hydrophilicity of cavities in proteins. Proteins: Structure, Function and Genetics 24, 433–438 (1996)
21. (accessed on 4th May 2008), http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html

Design of New Chemoinformatic Tools for the Analysis of Virtual Screening Studies: Application to Tubulin Inhibitors

Rafael Peláez[1,*], Roberto Therón[2], Carlos Armando García[2], José Luis López[1], and Manuel Medarde[1]

[1] Departamento de Química Farmacéutica-Facultad de Farmacia
[2] Departamento de Informática y Automática, Facultad de Ciencias, Campus Unamuno. Universidad de Salamanca, Spain
pelaez@usal.es

Abstract. Virtual screening (VS) experiments yield huge numbers of configurations (conformations plus rotations plus translations). In order to extract important structural information from such a complex database, new chemoinformatic tools are urgently needed. We have clustered and classified by means of "ad hoc" semiautomatic chemoinformatic tools the poses arising from docking experiments conducted on more than 700,000 compounds on tubulin. The results obtained in this way have been compared with those achieved by visual inspection protocols in an attempt to develop new useful tools.

Keywords: clustering, drug design, virtual screening, tubulin.

1 Introduction

The drug discovery process is very complex and demanding and usually requires cooperative interdisciplinary efforts.[1] Despite the great and steady methodological advances achieved through the years and the huge efforts devoted to this enterprise, more often than not the results are disappointing. The recent completion of the human genome project has not only unearthed a number of new possible drug targets but also has highlighted the need for better tools and techniques for the discovery and improvement of new drug candidates. [2-4] The development of these new tools will benefit from a deeper understanding of the drugs´ molecular targets as well as from more friendly and efficient computational tools. [5]

Many experimental and computational approaches are already accessible at the different stages of the drug design and development processes. [6-7] The available resources, the degree of knowledge of the problem at hand and personal bias often condition the choices. In the early stages of drug discovery, finding new active chemical structures is a very important goal. Both experimental and virtual (in silico) screenings can be used to explore chemical space, and often both strategies are combined. [8-10] Virtual screening (VS) is an in silico high-throughput screening (HTS) procedure which attempts to rank candidate molecules in descending order of likelihood of

* Corresponding author.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 189–196, 2009.
springerlink.com

biological activity, hence reducing the number of compounds for which very costly experimental evaluations are needed. [11]

Docking is one plausible choice when the 3D structure of the target is known. It consists in the positioning of the compounds in the target site (posing) and a ranking of the resulting complexes (scoring). [12] The strategy is functional even when the binding site is unknown or ill defined, and there have been many successful cases described. However, docking simulations model several complex phenomena with a minimal time investment, thus resulting in inconsistent performance. Attempts to alleviate the inherent limitations of the method, such as the use of different scoring functions to compensate for their simplicity, the incorporation of ligand flexibility, or the simulation of protein flexibility have met with different degrees of success. [13-15] During the final stages, the best ranked compounds are clustered and classified and have to be visually inspected by experienced chemists in order to confirm their goodness, before the final selection for the screening stages. This visual filtering is time consuming, expensive, and difficult to reproduce, as chemists with different backgrounds would make different choices. New chemoinformatic tools which can help the chemist in the selection of candidates are thus welcome. However, they are difficult to develop, as they have to somehow reproduce chemical insight. In this communication, we will report on our results on the development of new chemoinformatic tools aimed at facilitating such a process.

2 Docking Results

For the present discussion, we will take as a model the colchicine binding site of tubulin, an important anticancer target. Tubulin form the microtubules, which in turn take part in a number of pivotal cell functions such as division, motility, intracellular transport, and shape maintenance. Antimitotic drugs interfere with the dynamics of microtubules and are in clinical use as antitumour, antiprotozoal or antifungal chemotherapeutics.[16] The structure of the tubulin dimer complexed with several ligands (Figure 1) has been recently determined [17] and we have used it for high-throughput virtual screening. [18, 19]

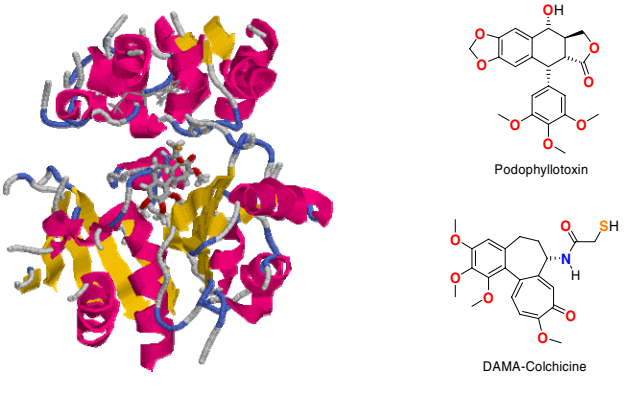
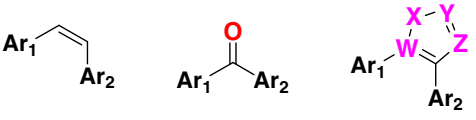


Fig. 1. Refined binding site of tubulin occupied by colchicine and chemical structures of podophyllotoxin and DAMA-colchicine

The dataset of hypothetical ligands docked consisted of virtual compounds from three different sources: a subset of 700.000 lead-like [20] molecules from the ZINC free electronic database [21], a subset of 990 nonreactive organic molecules from the ACD molecules employed by A. N. Jain in Surflex docking validation, [22] and a collection of more than 300 compounds synthesized in our laboratory and tested against tubulin.



Representative schematic structures of "in house" compounds

Even for a low number of poses for each of the compounds, the number of 3D structures generated during the virtual screening procedure is huge: 10 conformations for each of the more than 700,000 molecules make $7 \cdot 10^6$ conformations. The visual inspection of all the complexes is out question, and we have resorted to limit our initial analysis to the top scored 10000 compounds (1,000,000 conformations). Even for such a small part of the database, manual filtering is a daunting task. We therefore decided to attempt semiautomated means of comparison of the complexes (Figure 3). The clustering problem can be split in two parts: an initial clustering of the different poses (conformations, translations and rotations) for each ligand and a subsequent clustering of the selected poses of every ligand. The clustering results will be classified by comparing them with a chemical classification produced by one of us (R.P.) by comparing our ¨in house¨ dataset and the compounds from the ACD with the reference ligands: podophyllotoxin and DAMA-colchicine (for which the right pose is known from X-ray diffraction studies).

3 By Hand Clustering

A manual clustering of the top ranked poses yielded [23] eight distinct clusters of different sizes. According to chemical intuition, the docking programs generated a large number of top scoring poses with trimethoxyphenyl rings (or an equivalent moiety) in close proximity to those of colchicine and podophyllotoxin. These poses cluster in three different clusters, one of which was particularly enriched in compounds of the in house dataset that presented activity in tubulin polymerization assays. The aforementioned cluster also placed a second aromatic ring close to the tropolone of colchicine or to the methylenedioxyphenyl ring of podophyllotoxin. This results are in good agreement with known structure activity relationships for this family of compounds, which pinpoint that moiety as an essential structural requirement. The remaining two clusters place the second aromatic ring in quite different dispositions, suggesting that it is not possible to simultaneously occupy the two pharmacophoric points occupied by rings B and E of podophyllotoxin. Accordingly, and based on chemical intuition, these two clusters should correspond to less potent ligands, as is actually observed. The remaining clusters place the ligands in dispositions that bear no resemblance to those of the model ligands, and their interpretation is not so straightforward. A satisfactory semiautomatic clustering would be one that reasonably reproduces the described classification.

4 Semiautomated Atom Based Attempts to Cluster and Classify the Poses

We initially assumed that similar poses should put similar atoms in similar positions in space, as they would similarly react with the target. Therefore we generated a grid (we have tested point separations from 0,3 to 2 Å) that would sample the occupancy of the binding site by the poses. Using Insight macros, [24] the heavy (non hydrogen) atoms of the ligands we classified in three groups: aromatic atoms, aliphatic carbon atoms and aliphatic hetero-atoms. A bitstring was calculated for each ligand pose: for each grid point Insight macros [24] calculated if there was an atom –of each of the three types separatedly- within Van der Waals contact (Fig. 2). Therefore, each ligand pose was represented by a binary bitstring, of which three positions corresponded to each grid point. Similarity half-matrices were constructed by calculating the pairwise similarity (measured as similarity coefficients) of the bitstrings for every pose in the

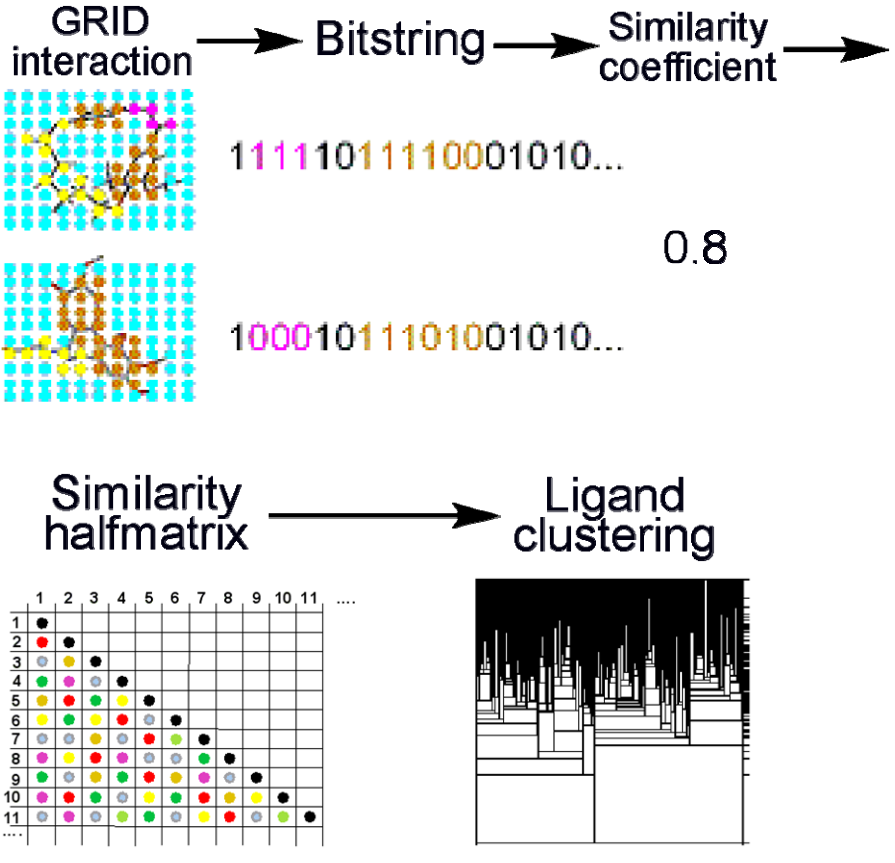


Fig. 2. Graphical summary of the semiautomated atom based clustering protocol

dataset to the rest of poses (Fig. 2). Initially, Tanimoto (alias Jaccard) coefficients were chosen for the similarity measures, since lack of occupancy was considered to convey no information, but later on different weighted coefficients were applied with similar results. [25]

The similarity half matrices thus generated were input to the XCluster software accompanying Macromodel® [26] and hierarchical similarity trees were generated. [27] The clustering procedure only led to chemically reasonable clusterings of the compounds when the number of compounds studied was small. For higher numbers of compounds, we always found more scattered results. The semiautomated protocol did not reproduce the visually generated clusters, as it was probably unable to capture the chemist viewpoint. Means for the incorporation of this bias in cluster generation seemed necessary to help the protocols reach a chemically reasonable outcome.

5 Semiautomated, Interactive, Grid Based Attempts to Cluster and Classify the Poses

Due to the unpleasant results obtained, we decided to follow a more interactive methodology which might allow us to cast the chemist knowledge onto the clustering results. The previous approach integrated the two clustering steps (clustering the poses for each ligand and clustering the ligands) in a single one. In order to promote interactivity, we decided to follow a two step procedure: first cluster the poses for each ligand and secong cluster the ligands.

The first step would help the chemist rationalize the docking results for each ligand and would be in itself an important help in the process of by hand clustering, as it would be a means of data reduction. Comparing different poses of the same ligand is often used in molecular modelling studies and is usually based on rmsd (root mean square deviation) calculations of the atomic coordinates in Cartesian space (Fig. 3). In addition to conformational differences, rotation and translation have to be considered in docking experiments, as opposed to other usual molecular modelling applications, such as conformational searching, were the conformations are superimposed by minimizing the rmsd, thus eliminating these contributions to the differences. Once again, a stepwise protocol in which a conformational comparison is first performed and then a comparison of the rotations and translations of the poses would be very convenient.

As previously mentioned, comparison of chemical entities with different numbers and types of atoms is a challenging task, and often chemists with different backgrounds make substantially different choices. A general procedure which automatically selects the important features of a drug – ligand complex to be considered in the pairwise comparison of ligands is thus difficult to implement. In order to carry this second step of the clustering problem, we have therefore decided to proceed in a staggered fashion: first, an automatic clustering should be attempted, which could be later on modified by the user.

In order to attempt an initial, unguided clustering, decisions have to be taken as to which regions of the binding site are more important to the recognition process. This

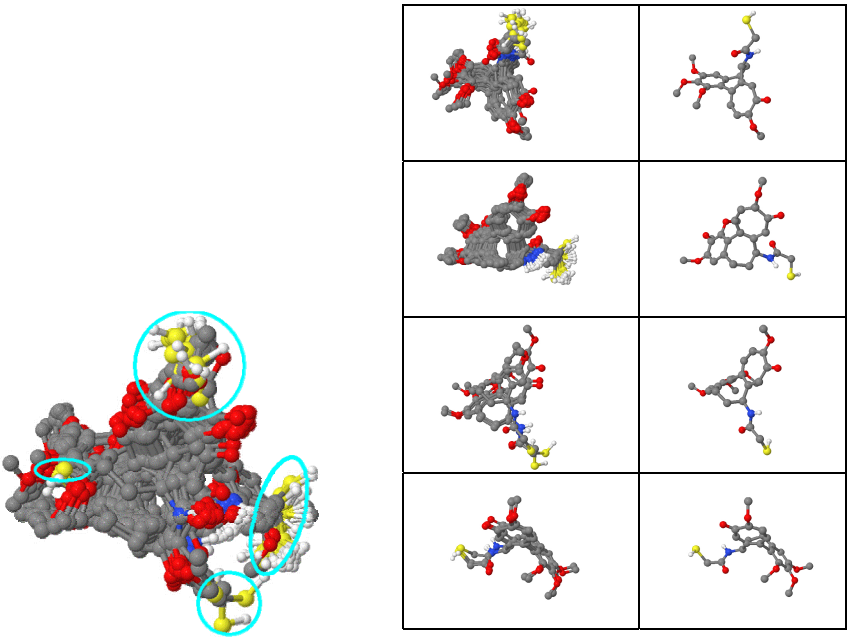


Fig. 3. Left: Superposition of 100 docked poses of colchicine. The blue lines indicate the location of the sulphur atoms (yellow spheres) in the four clusters found by hand. Right table: The four clusters with their specific representative molecules.

allows reduction of the data size without a loss of three dimensional resolution. We have therefore decided to start with the knowledge provided by the docking programs and select the highest rated (maxima) of the interaction energies. In order to do so, a docking program which operates based on a grid approximation was selected: AutoDock. [28] From the maps generated by AutoDock, zones of maximal and minimal contribution to the binding energy can be calculated. The occupation patterns of these privileged zones by the different poses of each ligand will provide the initial guesses for an automatic clustering.

At this point, user manipulation of the data is quite complicated due to the huge number of property values involved. In order to assist in the manipulation and filtering of the extraordinary quantity of information, new visual analytical tools. based on parallel coordinates will be applied. Towards this end, each pose of every compound is represented as a polyline that passes through parallel axis that gauges the properties previously selected. We have previously implemented a similar tool in a chemical database of natural products. [29]

6 Conclusions

New chemoinformatic tools have been designed for the analysis of the results of virtual screening protocols and their results have been compared to those of manual clustering. Grid based chemical comparison and manual clustering perform rather

inefficiently with large datasets and are very much time-consuming. With the help of visual analytical tools based on parallel coordinates chemical hindsight can be incorporated in the process of clustering, thus greatly reducing the amount of time invested. Moreover, chemical properties not easily derived from structure can be incorporated along the analysis, thus strengthening the correlation between clustering and biological activity and allowing for the derivation of more sounded structure activity relationships. These tools will be extremely useful in the selection of molecules as targets for chemical synthesis and assays.

Acknowledgments

We thank the MEC (Ref CTQ2004-00369/BQU), the Junta de Castilla y León (Refs. SA090A06, SAO30A06 and US21/06) and the EU (Structural Funds) for the financial support.

References

1. Burgers Medicinal Chemistry and Drug Discovery, 6th edn. John Wiley & Sons, Wiley-VCH Verlag GmbH & Co. (2003)
2. Lauss, M., Kriegner, A., Vierlinger, K., Noebammer, C.: Characterization of the drugged human genome. Pharmacogenomics 8, 1063–1073 (2007)
3. Overington, J.P., Al-Lazikani, B., Hopkins, A.L.: Opinion - How many drug targets are there? Nat. Rev. Drug Disc. 5, 993–996 (2006)
4. Paolini, G.V., Shapland, R.H.B., van Hoorn, W.P., Mason, J.S., Hopkins, A.L.: Global mapping of pharmacological space. Nat. Biotechnol. 24, 805–815 (2006)
5. Kubinyi, H.: Nat. Rev. Drug Disc. Drug research: myths, hype and reality 2, 665–669 (2003)
6. Ojima, I.: Modern Molecular Approaches to Drug Design and Discovery. Acc. Chem. Res. 41, 2–3 (2008)
7. Baxendale, I.R., Hayward, J.J., Ley, S.V., Tranmer, G.K.: Pharmaceutical Strategy and Innovation: An Academics Perspective. Chem. Med. Chem. 2, 268–288 (2007)
8. Ling, X.F.B.: High throughput screening informatics. Comb. Chem. High Throughput Screen 11, 249–257 (2008)
9. Carlson, T.J., Fisher, M.B.: Recent advances in high throughput screening for ADME properties. Comb. Chem. High Throughput Screen 11, 258–264 (2008)
10. Diller, D.J.: The synergy between combinatorial chemistry and high-throughput screening. Curr. Opin. Drug Discov. Devel. 11, 346–355 (2008)
11. Soichet, B.K.: Virtual screening of chemical libraries. Nature 432, 862–865 (2004)
12. Leach, A.R., Shoichet, B.K., Peishoff, C.E.: Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. J. Med. Chem. 49, 5851–5855 (2006)
13. Warren, G.L., Andrews, C.W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., Tedesco, G., Wall, I.D., Woolven, J.M., Peishoff, C.E., Head, M.S.: A Critical Assessment of Docking Programs and Scoring Functions. J. Med. Chem. 49, 5912–5931 (2006)
14. Carlson, H.A., McCammon, J.A.: Accommodating protein flexibility in computational drug design. Mol. Pharm. 57, 213–218 (2000)

15. Sotriffer, C.A., Dramburg, I.: In Situ Cross-Docking To Simultaneously Address Multiple Targets. J. Med. Chem. 48, 3122–3123 (2005)
16. Jordan, M.A., Wilson, L.: Microtubules as a target for anticancer drugs. Nat. Rev. Cancer 4, 253–265 (2004)
17. Ravelli, R.B., Gigant, B., Curmi, P.A., Jourdain, I., Lachkar, S., Sobel, A., Knossow, M.: Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. Nature 428, 198–202 (2004)
18. Alvarez, C., Alvarez, R., Corchete, P., Pérez-Melero, C., Pelaez, R.: Medarde. M. Bioorg. Med. Chem. 18 (2008), doi:10.1016/j.bmc.2008.04.054
19. Pelaez, R., López, J.L., Medarde, M.: Application of chemoinformatic tools for the analysis of virtual screening studies of tubulin inhibitors. In: Corchado, E., Corchado, J.M., Abraham, A. (eds.) Advances in Soft Computing, vol. 44, pp. 411–417. Springer, Heidelberg (2007)
20. Carr, R.A.E., Congreve, M., Murray, C.W., Rees, D.C.: Fragment-based lead discovery: leads by design. Drug Disc Dev. 14, 987–992 (2005)
21. Irwin, J.J., Shoichet, B.K.: ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. J. Chem. Inf. Model 45, 177–182 (2005)
22. Jain, A.N.: Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. J. Med. Chem. 46, 499–511 (2003)
23. Visualization of the superposed complexes was done with Jmol: http://www.jmol.org and with MarvinBeans 4.1.2 ChemAxon (2006), `http://www.chemaxon.com`
24. http://www.accelrys.com/
25. Kochev, N., Monev, V., Bangov, I.: Searching Chemical Structures. In: Chemoinformatics: A textbook, pp. 291–318. Wiley-VCH, Chichester (2003)
26. Mohamadi, F., Richards, N.G.J., Guida, W.C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T., Still, W.C.: MacroModel an Integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics. J. Comp. Chem. 11, 440–467 (1990)
27. Hart, D.R., Stark, P.: Pattern Classification. John Wiley and Sons, New York (2001)
28. Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comp. Chem. 19, 1639–1662 (1998)
29. López-Perez, J.L., Therón, R., del Olmo, E., Díaz, D.: NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. Bioinformatics 23, 3256–3257 (2007)

Multi-Objective Optimization of Biological Networks for Prediction of Intracellular Fluxes

José-Oscar H. Sendín, Antonio A. Alonso, and Julio R. Banga

Process Engineering Group
Instituto de Investigaciones Marinas (CSIC)
C/ Eduardo Cabello 6, 36208, Vigo, Spain
osendin@iim.csic.es, antonio@iim.csic.es, julio@iim.csic.es

Abstract. In this contribution, we face the problem of predicting intracellular fluxes using a multi-criteria optimization approach, i.e. the simultaneous optimization of two or more cellular functions. Based on Flux Balance Analysis, we calculate the Pareto set of optimal flux distributions in *E. coli* for three objectives: maximization of biomass and ATP, and minimization of intracellular fluxes. These solutions are able to predict flux distributions for different environmental conditions without requiring specific constraints, and improve previous published results. We thus illustrate the usefulness of multi-objective optimization for a better understanding of complex biological networks.

Keywords: Multi-objective optimization, Pareto front, Flux Balance Analysis.

1 Introduction

Intracellular fluxes in biochemical networks can be calculated *in silico* under the assumption that cellular systems operate in an optimal way with respect to a certain biological objective. Network capabilities and flux distributions have thus been predicted by using, for example, Metabolic Flux Balance Analysis (FBA), the fundamentals of which can be found in e.g. (Varma and Palsson 1994). FBA only requires the stoichiometric model of the network, but since the linear system of mass balance equations at steady-state is generally under-determined, appropriate cellular functions (objectives) must be defined, as well as other possible additional constraints, to find a unique solution. Successful applications of FBA include the prediction of *E. coli* metabolic capabilities (Edwards et al. 2001) and the genome-scale reconstruction of the metabolic network in *S. cerevisiae* (Forster et al. 2003).

In this context, a particularly interesting question which have been addressed recently in detail (Schuetz et al. 2007; Nielsen 2007) concerns the principles behind the optimal biochemical network operation, i.e.: "which are the criteria being optimized in these systems?" By far, the most common objective considered is the maximization of growth (or biomass yield), although other criteria, such as maximization of ATP yield (van Gulik and Heijnen 1995) or minimization of the overall intracellular flux (Bonarios et al. 1996), have been proposed for different systems and conditions.

Since neither we nor nature have a single goal, a more desirable and realistic approach is to consider the simultaneous optimization of two or more criteria, often conflicting. As a consequence, the solution will not be unique but instead this strategy

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 197–205, 2009.
springerlink.com

will result in a set of solutions representing the optimal trade-offs between the different objectives. Multi-objective (or multi-criteria) optimization is better able to cope with the complexity of models from systems biology (Handl et al. 2007), but few applications are found in literature in comparison with other scientific and engineering fields (Sendín et al. 2006).

In this work we face the solution of multi-objective optimization (MO) problems derived from FBA. By simultaneously optimizing several common cellular functions, the aim of this study is to test the capabilities of this approach for predicting intracellular fluxes independently from the environmental conditions and without imposing additional, case-dependent and potentially artificial constraints conditioning the final solution. After presenting the basic concepts and methods in MO, we will consider the central carbon metabolism in *Escherichia coli* as a case study to assess whether optimality principles can be generally applied.

2 Multi-Objective Flux Balance Analysis (MOFBA)

2.1 Problem Formulation and Basic Concepts

Assuming a biological network operating at steady-state, and if a stoichiometric model is available, the Multi-Objective Flux Balance Analysis problem can be stated as finding the flux distribution which optimizes simultaneously two or more objective functions subject to the mass balance equations:

$$\underset{\mathbf{v}}{\text{Max/Min}}\ \mathbf{Z} = \left[Z_1(\mathbf{v}) \quad Z_2(\mathbf{v}) \quad \ldots \quad Z_n(\mathbf{v}) \right]^{\mathrm{T}} \tag{1}$$

Subject to:

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0} \tag{2}$$

$$\mathbf{v}^L \leq \mathbf{v} \leq \mathbf{v}^U \tag{3}$$

Z is the vector of *n* objective functions (linear as well as non-linear); **S** is the (*m* x *r*) stoichiometric matrix, where *m* is the number of intracellular metabolites and *r* the number of reactions; **v** is the vector of *r* fluxes, with lower and upper bounds $\mathbf{v}^L$ and $\mathbf{v}^U$, respectively. Additional constraints can be imposed depending on the problem and the available experimental data and the knowledge about the system.

Simultaneous optimization of multiple objectives differs from traditional single-objective optimization in that if the objectives are in conflict with each other, there will not be a unique solution which optimizes simultaneously all of them. The key concept here is that of Pareto-optimal solution.

A point **v*** in the solution space is said to be Pareto-optimal if there does not exist another feasible point **v** such that $Z_i(\mathbf{v}) \leq Z_i(\mathbf{v}^*)$ for all $i=1,\ldots,n$ and $Z_j(\mathbf{v}) < Z_j(\mathbf{v}^*)$ for some *j*. In other words, **v*** is optimal in the sense that improvement in one objective can only be achieved by worsening one or more of the others. Thus, the solution of a MO problem is a family of potentially infinite points, none of which can be said to be better than another. This family is known as Pareto-optimal set or Pareto front.

2.2 Methods for Multi-Objective Optimization

Traditionally, multiple objectives are optimized simultaneously by defining a composite function combining different criteria. The most widely used approach consists in optimizing a weighted sum of the objectives, where each weight represents the relative importance of the associated objective. Within FBA, this type of utility functions has also been proposed, as e.g. maximization of ATP yield per flux unit (Dauner and Sauer 2001; Schuetz et al. 2007). However, this approach will yield only one optimal solution, overlooking the trade-off between the objectives.

In this work we have combined two well-known techniques for generating the complete Pareto-front (or at least a good representation of it):

- **ε-Constraint (EC):** This is also a common and intuitive method for solving a MO problem. In this approach, the original MO problem is transformed into a single-objective linear programming (LP) problem (if the objective functions and the constraints are linear) or a non-linear programming (NLP) problem by optimizing one of the objectives while the others are incorporated as inequality constraints. By changing the value of the parameter ε (i.e. the bounds on the objectives converted to constraints), different Pareto-optimal solutions can be obtained. Its main drawback is the difficulty to choose appropriate values for the parameters of the method to obtain a good picture of the Pareto front, so that no regions are over- or under- represented.
- **Normal Boundary Intersection (NBI):** This technique (Das and Dennis 1998) was developed to overcome the drawbacks of methods like the weighted sum approach in which it is difficult to obtain a complete representation of the Pareto-optimal set. Starting from the individual optima for each objective, NBI also converts the original MO problem into a set of LPs/NLPs in such a way that a systematic change in the method parameters generates an even spread of points on the Pareto front. Thus, the complete trade-off between the objectives can be captured by solving a lesser number of optimization problems. However, some regions of the Pareto surface can be missed in problems with more than two objectives.

It should be noted that global optimization (GO) solvers will be needed for both approaches if the associated single-objective NLPs are non-convex.

3 Case Study

Here we consider the central carbon metabolism in *Escherichia coli*, which has been studied in (Schuetz et al. 2007) to examine the predictive capacity of 11 linear and non-linear network objectives. The stoichiometric model consists of 98 reactions and 60 metabolites, and 10 split ratios R_i (i=1,…,10) at pivotal branch points were defined (Figure 1).

Taking as reference the above mentioned work, we address the problem in which three relevant cellular functions are optimized simultaneously:

$$\text{Find } \mathbf{v} \text{ to } \left\{ \begin{array}{l} \max Z_1(\mathbf{v}) = v_{Biomass} \\ \max Z_2(\mathbf{v}) = v_{ATP} \\ \min Z_3(\mathbf{v}) = \sum_{i=1}^{r} v_i^2 \end{array} \right\} \tag{4}$$

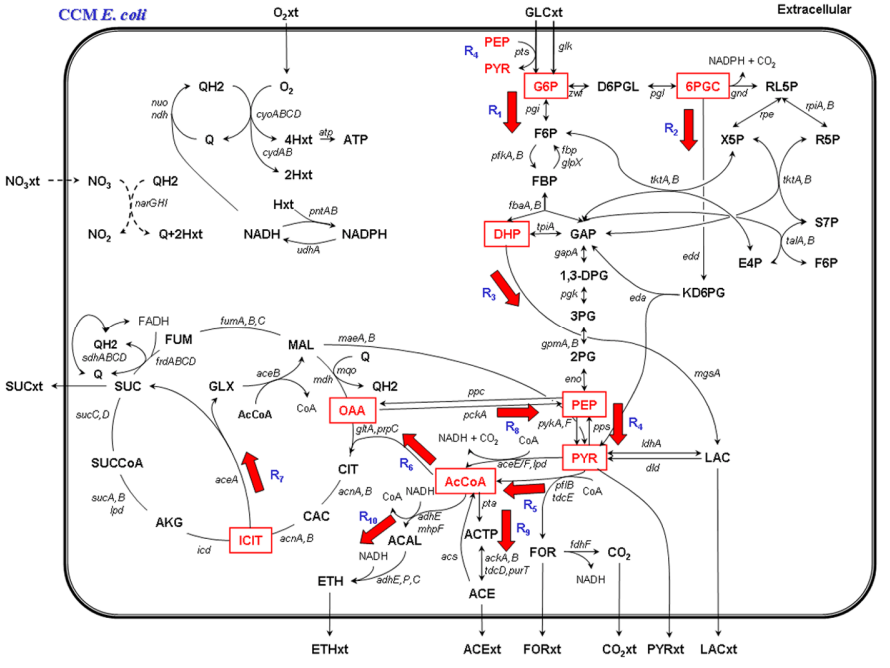


Fig. 1. Central carbon metabolism pathways in *Escherichia coli.* Red arrows represent the split ratios which describe the systemic degree of freedom of the network (for further explanation and abbreviations see Schuetz et al. 2007).

Table 1. Experimental flux split ratios (R) for the conditions considered

	Aerobic				Anaerobic
	Batch	Continuous C-limited	Continuous C-limited	Continuous N-limited	Batch NO_3 resp.
$v_{Biomass}$	8.3 mM/g·h	5.0 mM/g·h	7.0 mM/g·h	4.0 mM/g·h	1.77 mM/g·h
	ExpC1	ExpC2	ExpC3	ExpC4	ExpC5
R1	0.70 ± 0.02	0.69 ± 0.12	0.64 ± 0.05	0.96 ± 0.14	0.82 ± 0.02
R2	0.13 ± 0.06	0.23 ± 0.20	0.19 ± 0.11	0.00 ± 0.05	0.00 ± 0.05
R3	0.00 ± 0.05	0.00 ± 0.05	0.00 ± 0.05	0.00 ± 0.05	0.00 ± 0.05
R4	0.78 ± 0.02	0.84 ± 0.14	0.70 ± 0.06	0.72 ± 0.10	0.96 ± 0.02
R5	0.81 ± 0.03	0.91 ± 0.21	0.84 ± 0.14	0.90 ± 0.15	0.96 ± 0.02
R6	0.24 ± 0.02	0.64 ± 0.13	0.85 + 0.09	0.50 ± 0.06	0.02 ± 0.01
R7	0.00 ± 0.05	0.46 ± 0.13	0.00 ± 0.05	0.00 ± 0.05	0.00 ± 0.05
R8	0.00 ± 0.05	0.35 ± 0.08	0.12 ± 0.03	0.01 ± 0.01	0.00 ± 0.05
R9	0.58 ± 0.03	0.00 ± 0.05	0.00 ± 0.05	0.04 ± 0.01	0.65 ± 0.01
R10	0.00 ± 0.05	0.00 ± 0.05	0.00 ± 0.05	0.00 ± 0.05	0.30 ± 0.02

subject to the mass balance equations and the upper and lower bounds on fluxes (Eqs. 2-3). No additional constraints are imposed. It should be noted that objective functions Z_1 and Z_2 are linear, and the overall intracellular flux (Z_3) is non-linear, but convex.

Pareto-optimal solutions obtained with a combination of the methods described above will be compared with experimental flux data from *E. coli* (Table 1) under five environmental conditions (oxygen or nitrate respiring batch cultures and aerobic chemostats). The overall agreement is quantified using a standardized Euclidean distance between the computed split ratios and the experimental ones.

4 Results and Discussion

4.1 Optimization Settings

The three-objective optimization problem defined above is solved using a combination of the ε-constraint technique and NBI. The solution strategy consists of the following steps:

1. Maximize $v_{Biomass}$ using LP
2. Choose different values bm_i for $v_{Biomass}$ in the range $[0, v_{Biomass}^{max}]$
3. For each value bm_i, the following bi-objective optimization problem is solved using NBI: maximization of ATP and minimization of the overall intracellular flux subject to the ε-constraint: $v_{Biomass} \geq bm_i$

The resulting NLPs from application of NBI are solved by means of a multi-start clustering algorithm, *GLOBALm* (Sendín et al. 2008). This is a global optimization method which can detect the potential existence of multiple optima (i.e. solutions with the same value of the objective function and different flux profiles). For the sake of comparison with the results reported in (Schuetz et al. 2007), we have made use of the solvers included in the MATLAB® Optimization Toolbox (The MathWorks, Inc.): *linprog* for the LPs and *fmincon* as local solver within *GLOBALm* for the NLPs.

4.2 Pareto-Optimal Sets

The resulting Pareto surfaces (interpolated) for both aerobic and anaerobic conditions are showed in Figures 2 and 3, respectively. The trade-off between ATP yield and the overall intracellular flux is also depicted for each one of the biomass fluxes corresponding to the experimental conditions. Both Pareto-optimal sets obtained using the hybrid approach εConstraint-NBI are represented in Figure 4.

From inspection of these figures is clearly evident the existing conflict between ATP production and the overall intracellular flux for a given biomass flux. Maximum ATP yields (higher in the aerobic case) are achieved at the expense of an increase in the enzyme usage and with low growth rates. On the other side, biomass can be maximized while maintaining the overall intracellular flux at low levels. The cost to pay in this case is a decrease in the ATP yield.

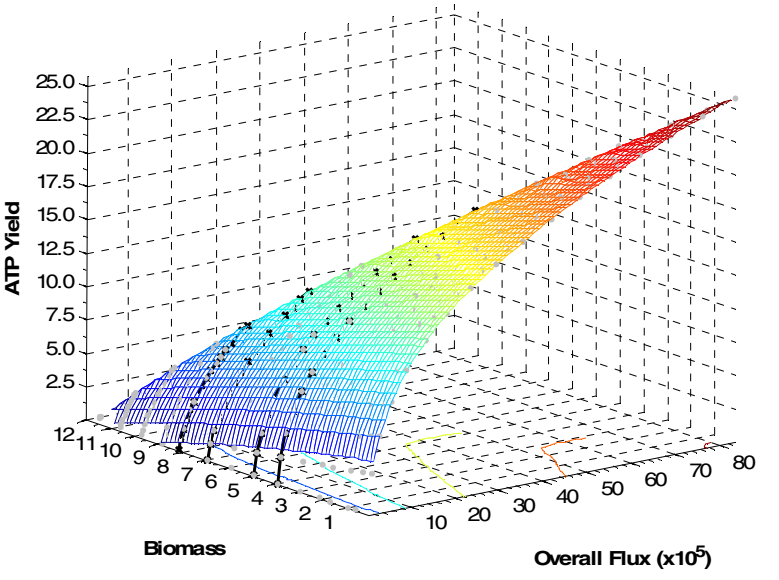


Fig. 2. Pareto front in aerobic conditions

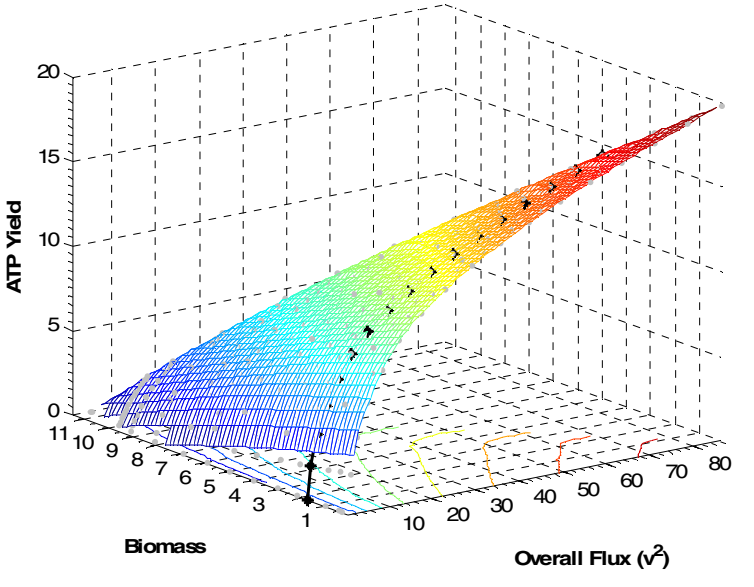


Fig. 3. Pareto front in anaerobic conditions

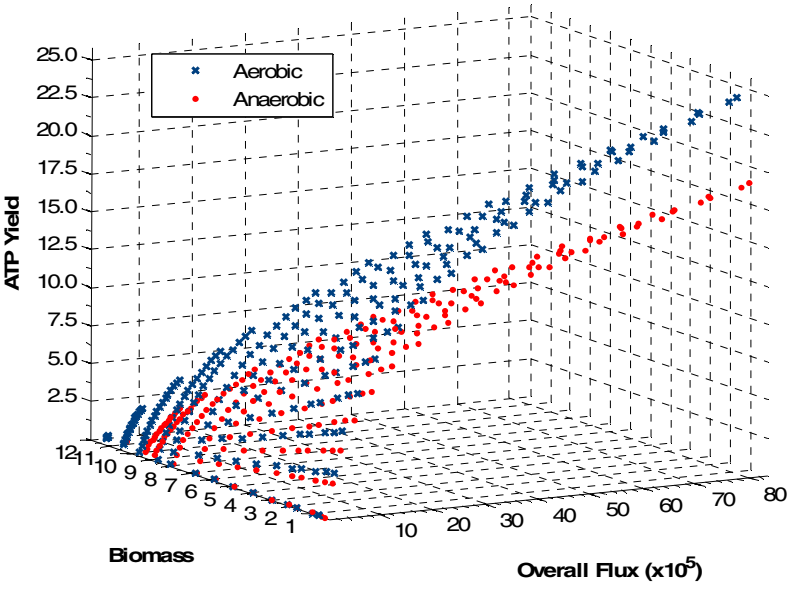


Fig. 4. Comparison of Pareto-optimal sets

4.3 Analysis of Solutions

Split ratios for each Pareto solution are compared with the experimental data, selecting those which yield the closest flux predictions (Table 2).

Table 2. Selected Pareto-optimal points

	ExpC1 A	**ExpC2** B	**ExpC3** C	**ExpC4** D	**ExpC5** E
$v_{Biomass}$	8.3	5.0	11.0	7.0	1.75
R1	0.74	0.98	0.64	0.98	0.55
R2	0.42	0.00	0.09	0.00	1.00
R3	0.00	0.00	0.0	0.00	0.00
R4	0.80	0.98	0.72	0.85	0.97
R5	0.81	0.92	0.70	0.86	0.52
R6	0.31	0.79	0.59	0.77	0.01
R7	0.00	0.12	0.0	0.00	0.00
R8	0.00	0.00	0.0	0.00	0.00
R9	0.50	0.04	0.05	0.09	0.74
R10	0.00	0.00	0.0	0.00	0.22

For the continuous cultures, the best predictions were found in Schuetz et al. (2007) when maximizing biomass or ATP yield coupled with several constraints. Somewhat similar flux distributions were obtained here, but we want to stress the fact that no additional, case-specific, constraints were imposed. For example, solution B (for C-limited continuous cultures) is similar to that resulting from maximization of ATP subject to an overproduction of 35% of NADPH relative to the NADPH requirement for biomass production, and the flux profile C maximizes biomass while satisfying a constraint on intracellular fluxes (limited to a 200% of the glucose uptake rate), and an upper bound on the oxygen uptake of 150% of the glucose uptake. For N-limited continuous cultures, point D also improves the prediction obtained when only one single objective is considered (with or without additional constraints).

5 Conclusions

In this work we have addressed the question of whether intracellular fluxes can be predicted considering optimality principles. The assumption here is that fluxes are distributed to optimize not only one single cellular function but several objectives simultaneously (multi-objective optimization).

In general terms, Pareto-optimal flux distributions improve the best predictions obtained with traditional FBA using different combinations of objective functions and constraints. The advantage of the multi-objective approach is that no additional, case-specific, constraints are needed, and it can be a powerful tool for a better understanding of the factors that influence the metabolic flux.

Acknowledgments

This work has been supported by EU project BaSysBio LSHG-CT-2006-037469.

References

Bonarios, H.P.J., Hatzimanikatis, V., Meesters, K.P.H., de Gooijer, C.D., Schmid, G., Tramper, J.: Metabolic flux analysis of hybridoma cells in different culture media using mass balances. Biotechnology Bioengineering 50, 299–318 (1996)

Das, I., Dennis, J.E.: Normal Boundary Intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. SIAM J. Optimization 8, 631–657 (1998)

Dauner, M., Sauer, U.: Stoichiometric growth model for riboflavin-producing Bacillus subtilis. Biotechnology Bioengineering 76, 132–143 (2001)

Edwards, J.S., Ibarra, R.U., Palsson, B.O.: In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology 19(2), 125–130 (2001)

Forster, J., Famili, I., Fu, P., Palsson, B.O., Nielssen, J.: Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Research 13(2), 244–253 (2003)

Handl, J., Kell, D.B., Knowles, J.: Multiobjective optimization in bioinformatics and computational biology. IEEE-ACM Transactions on Computational Biology and Bioinformatics 4(2), 279–292 (2007)

Nielsen, J.: Principles of optimal metabolic network operation. Molecular Systems Biology 3, 2 (2007)

Schuetz, R., Kuepfer, L., Sauer, U.: Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Molecular Systems Biology 3, 119 (2007)

Sendín, J.O.H., Vera, J., Torres, N.V., Banga, J.R.: Model-based optimization of biochemical systems using multiple objectives: a comparison of several solution strategies. Mathematical and Computer Modelling of Dynamical Systems 12(5), 469–487 (2006)

Sendín, J.O.H., Banga, J.R., Csendes, T.: Extensions of a multistart clustering algorithm for constrained global optimization problems. Industrial & Engineering Chemistry Research (accepted for publication, 2008)

van Gulik, W.M., Heijnen, J.J.: A metabolic network stoichiometry analysis of microbial growth and product formation. Biotechnology Bioengineering 48, 681–698 (1995)

Varma, A., Palsson, B.O.: Metabolic flux balancing: Basic concepts, scientific and practical us. Bio/Technology 12(10), 994–998 (1994)

SimSearch: A New Variant of Dynamic Programming Based on Distance Series for Optimal and Near-Optimal Similarity Discovery in Biological Sequences

Sérgio A.D. Deusdado[1] and Paulo M.M. Carvalho[2]

[1] ESA, Polytechnic Institute of Bragança, 5300, Bragança, Portugal
sergiod@ipb.pt
[2] Department of Informatics, School of Engineering, University of Minho, 4710, Braga, Portugal
pmc@di.uminho.pt

Summary. In this paper, we propose SimSearch, an algorithm implementing a new variant of dynamic programming based on distance series for optimal and near-optimal similarity discovery in biological sequences. The initial phase of SimSearch is devoted to fulfil the binary similarity matrices by signalling the distances between occurrences of the same symbol. The scoring scheme is further applied, when analysed the maximal extension of the pattern. Employing bit parallelism to analyse the global similarity matrix's upper triangle, the new methodology searches the sequence(s) for all the exact and approximate patterns in regular or reverse order. The algorithm accepts parameterization to work with greater seeds for near-optimal results. Performance tests show significant efficiency improvement over traditional optimal methods based on dynamic programming. Comparing the new algorithm's efficiency against heuristic based methods, equalizing the required sensitivity, the proposed algorithm remains acceptable.

Keywords: Similarity discovery, dynamic programming, distance series.

1 Introduction

Dynamic programming (DP) is a mathematical technique widely used in multiple research fields providing optimal solutions for complex problems, mostly of combinatorial nature. The final and optimal solution is achieved by analysing the sub-problems recursively and their optimal solutions, combining and integrating the partial solutions. Bellman's seminal work [3] marks the beginning of DP automatization towards informatics, describing a class of algorithms to implement DP. Later on, and naturally, this technique has been adopted by bioinformatics.

In bioinformatics, DP is fundamentally used to discover sequence alignments, considering both local and global alignments [9]. This task involves basically the search for similarities, analysing the involved sequences and pointing out the correct correlated segments. Since biological homologies are commonly approximate, the similarities may contain an acceptable degree of deviation, corresponding to an admissible number of mismatches; thus this attribute increases the problem's complexity. Similarity evaluation is based on the "edit distance" concept. The edit distance corresponds to the minimum

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 206–216, 2009.
springerlink.com

number of operations required to convert a sequence into another using three edit operations to insert, delete or substitute symbols. In order to evaluate the correlation degree, a scoring scheme is necessary to assess the similar regions and, on the other hand, penalize deviations (mismatches, substitutions and gaps). The obtained scores are stored in a similarity or scoring matrix providing the basis for further analysis.

DP is called an exhaustive technique since it tests all possible combinations and provides 100% sensitivity. Heuristic based methodologies are evolving as alternative solutions to attain fast and efficient near-optimal similarity discovery. Searching the sequences with partial but reasonable sensitivity allows to achieve the majority of homologies, it is a tradeoff solution since heuristic based algorithms use a fraction of the processing time and resources required by optimal methods.

In this paper, we propose SimSearch, an algorithm implementing a new variant of dynamic programming based on distance series, for optimal and near-optimal similarity discovery in biological sequences. The use of distance series enables a simpler similarity matrix construction, as well as, it only requires binary digits to represent similarity regions. The new searching strategy does not use a scoring scheme to store scores in the matrix, instead, the scoring scheme is useful subsequently, during the matrices' analysis to detect the pattern's presence and extent. Backtrack analysis is used to identify continuities among the sub-results from adjacent matrices, eventually needed to compose wider similarity regions.

The performance tests carried out in our work, show significant efficiency improvement over traditional optimal methods based on dynamic programming.

2 Related Work

In the domain of bioinformatics two well-known methodologies, based on dynamic programming, were created and evolved in the last three decades to provide 100% sensibility and complete similarities discovery. The first use of the dynamic programming approach for global alignment of biological sequences was reported by Saul Needleman and Christian Wunsch in 1970 [13] and then, in [18] slightly modified by Sellers. In 1981, Smith and Waterman proposed a new algorithm [19] in order to solve the local alignment problem.

The comparison of sequences using dynamic programming is often a slow process due to the intensive computing required. To overcome this constraint new approaches based on heuristics were developed to obtain near-optimal solutions, reducing significantly the time required to perform the homology search and, concomitantly reducing the processing and memory requirements to complete the task. Representative heuristic-based algorithms are BLAST [1, 2, 6] and PatternHunter [11, 12].

The use of distance series to analyse biological sequences is not a novelty in bioinformatics, statistical analysis using distance series are common [7, 20]. However, this methodology is still underdeveloped as regards sequence comparison and pattern discovery.

Despite the trend of alignment applications development in recent years indicating a preponderance of heuristic-based solutions, optimal homology search is still a necessity in several bioinformatics applications, such as biological data compression, DNA linguistics study and others. Therefore, the motivation for the development of efficient

optimal homology search algorithms remains and SimSearch is a contribution in that direction.

2.1 Smith-Waterman Algorithm

When considering optimal alignment and dynamic programming, the most used algorithm to compute the optimal local alignment is the Smith-Waterman [19] with Gotoh's [4] improvements for handling multiple sized gap penalties. The two sequences to be compared, the query sequence and the database sequence, are defined as $Q=q_1, q_2 \ldots q_m$ and $D=d_1, d_2 \ldots d_n$. The length of the query sequence and database sequence are $m=|Q|$ and $n=|D|$, respectively. A scoring matrix $W(q_i, d_j)$ is defined for all residue pairs by applying the recurrence relation (2.3). Usually the weight $W(q_i, d_j) \leq 0$ when $q_i <> dj$ and $W(q_i, d_j) > 0$ when $q_i=d_j$. The penalty for starting a gap and continuing a gap are defined as G_{init} and G_{ext}, respectively.

The alignment scores ending with a gap along D and Q are Equation (2.1) and Equation (2.2), respectively.

$$E_{i,j} = \max \left\{ \begin{array}{c} E_{i,j-1} - G_{ext} \\ H_{i,j-1} - G_{init} \end{array} \right\} \quad (2.1)$$

$$F_{i,j} = \max \left\{ \begin{array}{c} F_{i-1,j} - G_{ext} \\ H_{i-1,j} - G_{init} \end{array} \right\} \quad (2.2)$$

$$H_{i,j} = \max \left\{ \begin{array}{c} 0 \\ E_{i,j} \\ F_{i,j} \\ H_{i-1,j-1} - W(q_i, d_j) \end{array} \right\} \quad (2.3)$$

The values for H_{ij}, E_{ij} and F_{ij} are equal to 0 when $i < 1$ or $j < 1$.

By assigning scores for matches or substitutions and insertions/deletions, the comparison of each pair of characters is weighted into a matrix by calculation of every possible path for a given cell. In any matrix cell the value represents the score of the optimal alignment ending at these coordinates and the matrix reports the highest scoring alignment as the optimal alignment.

The Smith-Waterman algorithm runs in quadratic time and requires quadratic space: $O(nm)$ to compare sequences of lengths n and m.

2.2 BLAST Algorithm

The first stage of BLAST [1, 2, 6] - and many other homology search tools - involves identifying hits: short, high-scoring matches between the query sequence and the sequences from the collection being searched. The definition of a hit and how they are identified differs between protein and nucleotide searches, mainly because of the difference in alphabet sizes. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence (seeds), which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighbourhood word score threshold [1]. These initial neighbourhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extensions of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its

maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters *W*, *T*, and *X* determine the sensitivity and speed of the alignment. The BLAST program uses BLOSUM62 scoring matrix [5] alignments and, by default, a word length (*W*) of 11.

2.3 PatternHunter Algorithm

The novel introduction of spaced seeds in the filtration stage of sequence comparison by Ma et al. [12] has greatly increased the sensitivity of homology search without compromising the speed of search. This is the underlying idea of PatternHunter I [12] and II [11], a new generation of general purpose homology search tools, considered prominent solutions in similarity discovery in biological data. PatternHunter is a heuristics-based algorithm, highly efficient and sensitive due to the use of spaced seeds. Spaced seeds are strategically designed gapped *n-grams* aiming to identify similarity regions in biological sequences. In PatternHunter, the residues of a seed are interleaved with "don't care" filler residues in order to allow matches to sub-strings in the query sequences that do not include all the residues in the seed one after the other. Further analysis is used to extend the seed in both directions in an effort to obtain longer homologous subsequences. Multiseed filtration was a natural evolution of PatternHunter, achieving improved sensitivity.

3 The New Algorithm

SimSearch is originally an optimal similarity discovery algorithm as it performs an exhaustive search for repeats. However, it includes a filtering strategy which consists of using a minimum length seed to accelerate the discovery of high scoring similarity regions. The SimSearch seeds are, in fact, oriented to locate pairs of patterns already found in the sequence(s) contrarily to the seed concept of other algorithms such as BLAST or PatternHunter. In this way, the proposed algorithm searches for existing patterns already registered in the similarity matrices.

The similarity identification method used by SimSearch is based on the analysis of distance series of symbols and its correlation. Basically, regarding genomic sequences, each one of the four bases *(a,c,g,t)* occurs in the sequence frequently but not periodically, so collecting the distance (in bases) between occurrences can be useful to discover similarity. In fact, if *n* consecutive symbols appear in the sequence equidistant to a next occurrence then it is safe to affirm that a similarity of length *n* is present.

Initially, the sequence is analysed to determine and list the distance from the last occurrence of each symbol. Using a linked list for each symbol, the first occurrence of each symbol will constitute the head of the linked list, the node stores a value corresponding to the number of symbols preceding it in the sequence. The next occurrences will form the next nodes, containing information about the distance from the last symbol's occurrence. The tail of the list will correspond to the last occurrence of the symbol in the sequence. Formally, the distance series are composed by a number of terms equal to the number of identical symbol's occurrences; each term is

a number representing the distance in symbols to reach the next occurrence of an identical symbol, except for the first term whose value corresponds to its ordinal position in the sequence.

SimSearch can be adapted to execute different sequences' comparisons or similarity discover within a sequence. SimSearch was initially conceived to support a biological data compression application, thus to find out repetitions within the sequence. In this case, we only have the query sequence, the database sequence is used, for instance, when searching palindromes - the patterns present in the reverse complementary sequence. To compare two sequences, the SimSearch linked lists must register the distance to the next occurrence of the base *b* in the database sequence to the same base *b* in the query sequence.

3.1 The Similarity Matrices

Each similarity matrix is filled out writing the symbol "1" in the matrix lines present in the linked list, corresponding to the symbol in each column. In this manner, for each symbol of the query sequence only the future occurrences are signalled. The inherent simplicity of the process provides great efficiency, allowing the analysis of large sequences using moderate processing time and computing resources.

Going into details, and providing a practical example, lets consider a genomic sequence *x*="*tatccgcattatgcgata*", with length *m*=18. Table 1 contains the resulting similarity matrix, showing the patterns corresponding to successive 1s. A pattern initiated at *x[0]*, of length 6 and containing a mismatch was highlighted, the respective repetition occurs 9 (line number) symbols ahead at *x[9]*.

Table 1. The binary similarity matrix based on distance series

	T	A	T	C	C	G	C	A	T	T	A	T	G	C	G	A	T	A
1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
2	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1		
3	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0			
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
5	0	0	0	0	0	0	0	0	0	0	1	1	0					
6	0	1	1	0	0	0	0	0	0	0	0	0						
7	0	0	1	0	0	1	1	0	0	1	1	0						
8	1	0	0	0	0	0	0	1	1	0	0							
9	1	1	1	0	1	1	0	0	0	0								
10	0	0	0	1	0	0	0	1	0									
11	1	0	0	0	0	0	0	0										
12	0	0	0	0	0	0	0											
13	0	0	0	0	0	0												
14	0	1	1	0	0													
15	0	0	0	0														
16	1	1	0															
17	0	0																
18	0																	

The number of 1s in the similarity matrix can be determined using the following theorem (Theorem 3.1).

Theorem 3.1
Being $S(n)=1+2+3+4+5+...+n$, with $n \in \mathbb{N}$, described as the "sumtorial" of n, considering the expression $S(n)=\frac{n(n+1)}{2}$ as a valid method to calculate $S(n)$ and having n_1 as the sum of symbol "A" occurrences, n_2 as the sum of symbol "C" occurrences, n_3 as the sum of symbol "G" occurrences and n_4 as the sum of symbol "T" occurrences, the expression to calculate the number of 1s to be written in the similarity matrix will be:

$$\textbf{\textit{Number of 1s} = } S(n_1-1) + S(n_2-1) + S(n_3-1) + S(n_4-1)$$

Applying the theorem to the segment x="*tatccgcattatgcgata*", with $n_1=5$, $n_2=4$, $n_3=3$ e $n_4=6$, the total number of cells signalled with 1 will be: $S(4)+ S(3)+ S(2)+ S(5)=10+6+3+15=34$.

Analysing the time complexity involved in the process of filling out the matrix, the worst case occurs when the sequence is uniform (all the bases are equal), and, in this case, the complexity can be expressed as $O(S(n-1))$, being n the length of the sequence. The average case, also the most likely to occur, corresponds to a balanced distribution of all the four genomic symbols, thus the time complexity is $O(4S((n/4)-1))$.

SimSearch operates by dividing the sequence to examine in manageable subsequences, specifically segments of dimension d=10.000 symbols by default. Facing the impossibility of representing here the full SimSearch subtable, Table 1 shows a division of the whole matrix in 3*3 sub-matrices. The number of matrices to process is also calculated by the expression in Theorem 3.1. Being n the length of the sequence and d the dimension of the matrices, the number of matrices to process will be S(n/d).

The next step is devoted to filter the similarity occurrences, avoiding irrelevant patterns. The "irrelevant" definition depends on the assumed scoring scheme and the resulting distance edition cost.

3.2 Similarity Discovery

The similarity matrices include all the present similarities, i.e., 100% sensitivity at this stage. To uncover the similar segments requires the analysis of the similarity matrices in order to identify different kinds of successions of 1s in the matrix. The shape of the successions indicates the quality of the patterns. In fact, the proposed methodologies allow the identification of exact, approximate, reverse, or palindrome patterns.

The discovery of patterns begins with the detection of seeds. Considering a seed of length W=9, the algorithm analyses the matrices in several ways suspending the search when a seed is located (in the example, nine successive 1s). A local analysis is carried out to examine the maximum extent of the pattern. The proposed algorithm uses bitwise operations to locate the seeds quickly; an adaptation of the algorithm SBNDM [15] was also included for fast seed matching. SimSearch seeds are limited to 32 characters due to restrictions of bit-parallelism operations.

3.3 The Scoring Scheme

SimSearch uses a scoring scheme to analyse the similarity matrices, although it is not registered in the matrices. Whenever a seed is detected, the scoring scheme is used

attempting to extend the pattern in both sides up to a pre-defined threshold. This means that the attempt to discover a larger pattern allows a certain number of mismatches. SimSearch's default scoring scheme is: *Match=1 point; Mismatch=-1 point; Open a gap=-5 points.* However, the user may redefine these values.

The accumulation of successive negative score, exceeding a defined threshold value, will imply the end of the pattern analysis. On the other hand, only patterns with a minimum score are considered.

For certain applications, a minimum pattern length (or score, in a different perspective) is required, so SimSearch includes a parameter to set this value.

3.4 Exact Repetitions and Overlapped Patterns

As referred above, the exact repetitions are identified by consecutive successions of 1s in the lines of the similarity matrices. An important disruption that affects pattern discovery algorithms corresponds to the existence of overlapped patterns within a larger pattern. This normally implies redundant processing as overlapped patterns are not relevant to functional genomics as they are to stringology.

The proposed algorithm, through its similarity matrices, keeps record of all the patterns, including the overlapped ones. For instance, the pattern "*aaaaa*" contains three occurrences of the pattern "*aaa*". Similarily, the pattern "*ctctctc*" contains three occurrences of the pattern "*ctc*". In the analysis stage, SimSearch includes mechanisms to avoid redundant processing regarding overlapped sub-patterns.

Table 2 contains the resulting similarity matrix considering the sequence "*atcatcatc*". For simplicity reasons, the matrix is only filled with the relevant 1s.

Analysing the similarity matrix (see Table 2), two patterns are visible:

- the first corresponds to "*atcatc*", present both at 1st and 4th bases. However, there are overlapped sub-patterns within it.
- the second pattern corresponds to a sub-pattern of the first pattern, where the segment "*atc*" is repeated at 1st and 7th bases, thus there is no significant discovery as the first pattern covers the same region.

Tables 2 and 3. Examples of redundant patterns

	A	T	C	A	T	C	A	T	C
1									
2									
3	1	1	1	1	1	1			
4									
5									
6	1	1	1						
7									
8									

	A	A	A	A	T	A	A	A	A
1	1	1	1			1	1	1	
2	1	1		1		1	1		
3	1		1	1		1			
4		1	1	1					
5	1	1	1	1					
6	1	1	1						
7	1	1							
8	1								

Multiple overlapped patterns cause a dense region of 1s in the similarity matrix as shown in Table 3, considering the sequence "*aaaataaaa*". The larger exact repeat, "*aaaa*", occurs at 1st and 5th bases. Several minor overlapped patterns occur within the larger ones, for instance, the pattern "*aa*" occurs six times.

In order to overcome the redundancy inherent to overlapped patterns, SimSearch protects itself from over-searching. This is accomplished calculating an exclusion region based on the upper triangle limited by the diagonal that encloses the larger pattern (see the triangle in Table 3). In this example, several minor patterns were discovered before the larger pattern. The discovery of the larger pattern will dismiss the minor ones and avoid the subsequent search for irrelevant patterns within the larger one.

3.5 Reverse Repetitions

Basically a reverse pattern is a symmetric pattern. Considering the sequence "*attgcgtta*", the pattern "*attg*" occurs in the reverse form at the end of the sequence. Table 4, representing the similarity matrix originated by the sequence "*attgcgtta*", shows that the reverse pattern is also registered in the matrix, but not in a horizontal line as in the exact pattern occurrences. The reverse patterns are registered in the diagonal lines, not the regular diagonal (45º) but in the 67,5º diagonal.

Table 4. A reverse repetition detected in the sequence "*attgcgtta*"

	A	T	T	G	C	G	T	T	A
1		1					1		
2				1					
3									
4			1						
5		1	1						
6		1							
7									
8	1								

3.6 Approximate Repetitions

In genomics, approximate repetitions are naturally more abundant than exact repetitions, the explanation lies on the results of genomic maintenance and evolutionary mechanisms [17]. The study of approximate repeats is important in functional genomics and several bioinformatics tools are available exclusively focused on this subject [8, 10]. The SimSearch similarity matrices also record the pattern interruptions (see Tables 5 and 6).

Tables 5 and 6. Approximate patterns represented in the similarity matrices

	A	G	A	C	T	A	A	A	C
1						1	1		
2	1					1			
3			1						
4			1						
5	1	0	1	1					
6	1								
7	1								
8									

	A	A	G	G	A	A	T	G	G	A
1	1		1		1			1		
2										
3		1								
4	1	1		1						
5	1		1	1	1					
6			1							
7										
8		1								
9	1									

Substitutions

The proposed algorithm identifies substitutions in a pattern if a succession of 1s, in the same line, contains interspersed 0s. Analysing the sequence "*agactaaac*" in Table 5, an approximate pattern containing one substitution is present, specifically, the segments "*agac*" and "*aaac*" only differ on the second base.

Insertions/Deletions (Indels)

Approximate repeats may also include rearrangements such as insertions or deletions. In fact, the symmetry property implies an insertion for each deletion and vice-versa, i.e., if a certain string needs an insertion to equalize a second string then the equalization may also be operated with a deletion in the second string. The word indel is used in these cases.

For instance, when analysing the sequence "*aaggaatgga*", the pattern "*aagga*" occurs twice depending on a deletion of the seventh base. From other perspective, the pattern "*aatgga*" occurs twice depending on an insertion ("*t*") in the third base. The SimSearch similarity matrices reflect these situations as a succesion of 1s whose continuity may be extended in adjacent lines (see Table 6). If consecutive indels occur (gap) then the distance between lines corresponds to the extent of the gap.

4 Results and Discussion

The performance analysis of sequence alignment is necessary to assess the efficiency of the different methodologies [16]. In order to evaluate and compare the performance of the new algorithm three different competitors were chosen considering their importance, methodology and efficiency. The first choice was SSearch [14] (version 35), a classical dynamic programming algorithm used to achieve optimal local alignments. SSearch uses Pearson's implementation of the method of Smith and Waterman. The second choice was the BLAST tool (version 2.2.18), the most famous and widely used application in bioinformatics to execute local and global alignments. The third choice was a state-of-the-art and top performing tool, PatternHunter (version 2).

SimSearch's default settings are, mainly, similarity matrices of 10.000*10.000 cells, seed length *W=11* and local alignments with scores no less than *16*. The scoring scheme is the one defined in section 3.3. Other contenders use, when applicable, the same settings to provide equal conditions. SimSearch was tested for optimal homology search using *W=3*. PatternHunter was tested using 2 and 8 seeds for different sensitivity comparisons.

A prototype of SimSearch was developed, implemented in C language, and compiled with best optimizations using *gcc*. Performance tests were executed using a system based on an Intel Pentium IV - 3,4 GHz - 512KB cache - 1GB DDR-RAM.

Two genomic sequences were tested separately for similarity regions dicovery: a human gene (*humghcsa*) with ~65 Kbases and an entire genome (*e. coli*) with ~4,6 Mbases. The assessed processing times are included in Table 7.

SimSearch detects assertively all the patterns present in a genomic sequence, PatternHunter and especially BLAST miss some patterns due to heuristics limitations. SimSearch overperforms clearly traditional DP algorithms being comparable with heuristics-based methodologies when the required sensitivity is near maximal.

Table 7. Performance vs. sensitivity comparison

	Execution times					
	SimSearch (optimal)	SimSearch (near-optimal)	SSearch	BLASTn	PHunter (2 seeds)	PHunter (8 seeds)
humghcsa	24 seconds	14 seconds	7,5 hours	4 seconds	2 seconds	6 seconds
e. coli	1,22 days	17,67 hours	5 years*	52 seconds	14 seconds	45 seconds

*estimation

5 Conclusions

In this paper we have proposed SimSearch, a new genomic-oriented homology search algorithm. Similarity discovery within a sequence or among sequences is an intensive and very time-consuming computational task. Optimal alignment solutions are, even today, computationally prohibitive as large genomic sequences analysis can take years using a conventional DP approach. The proposed algorithm is an exhaustive search algorithm, based on the analysis of similarity matrices obtained registering properly each symbol's distance series. The new algorithm follows a dynamic programming logic and achieves an optimal solution with 100% sensitivity. To improve processing time, SimSearch includes a filtering methodology centering attention in high score segments searched using a fast exact pattern-matching module. SimSearch is incomparably faster than the Smith-Waterman algorithm, the most used optimal local alignment solution based on DP. Compared with heuristics-based solutions SimSearch is, obviously, slower, but its competitiveness can be increased if parameterized to near-optimal search.

Acknowledgements. This work has been partially supported by PRODEP.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990)
2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997)
3. Bellman, R.E.: Dynamic Programming. Princeton University Press, Princeton (1957)
4. Gotoh, O.: An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708 (1982)
5. Henikoff, Henikoff.: Amino Acid Substitution Matrices from Protein Blocks. Natl. Acad. Sci. USA 89, 10915 (1989)
6. Huang, X., Miller, W.: A time-eficient, linear-space local similarity algorithm. Adv. Appl. Math. 12, 337–357 (1991)
7. José, M.V., Govezensky, T., Bobadilla, J.R.: Statistical properties of DNA sequences revisited: the role of inverse bilateral symmetry in bacterial chromosomes. Physica A: Statistical Mechanics and its Applications 351, 477–498 (2005)

8. Kolpakov, R., Bana, G., Kucherov, G.: mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. 31, 3672–3678 (2003)
9. Kruskal, J.B.: An overview of sequence comparison. Addison Wesley, Reading (1983)
10. Lefebvre, A., Lecroq, T., Dauchel, H., Alexandre, J.: FORRepeats: detects repeats on entire chromosomes and between genomes. Bioinformatics 19, 319–326 (2002)
11. Li, M., Ma, B., Kisman, D., Tromp, J.: PatternHunter II: Highly Sensitive and Fast Homology Search. J. Bioinform. Comput. Biol. 2, 417–439 (2004)
12. Ma, B., Tromp, J., Li, M.: Pattern Hunter: fast and more sensitive homology search. Bioinformatics 18, 440–445 (2002)
13. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443–453 (1970)
14. Pearson, W.R.: Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 11, 635–650 (1991)
15. Peltola, H., Tarhio, J.: Alternative Algorithms for Bit-Parallel String Matching. In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L. (eds.) SPIRE 2003. LNCS, vol. 2857, pp. 80–93. Springer, Heidelberg (2003)
16. Sanchez, F., Salami, E., Ramirez, A., Valero, M.: Performance Analysis of Sequence Alignment Applications. In: IEEE International Symposium on Workload Characterization, pp. 51–60 (2006)
17. Schmidt, T., Heslop-Harrison, J.S.: Genomes genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci. 3, 195–199 (1998)
18. Sellers, P.H.: On the theory and computation of evolutionary distances. SIAM J. Appl. Math. 26, 787–793 (1974)
19. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197 (1981)
20. Teodorescu, H.-N., Fira, L.-I.: Analysis of the predictability of time series obtained from genomic sequences by using several predictors. Journal of Intelligent and Fuzzy Systems 19, 51–63 (2008)

Tuning Parameters of Evolutionary Algorithms Using ROC Analysis

Lino Costa, Ana Cristina Braga, and Pedro Oliveira

Department of Production and Systems Engineering
University of Minho
Campus de Gualtar, 4710-057 Braga, Portugal
{lac,acb,pno}@dps.uminho.pt

Summary. Evolutionary algorithms mimic the natural evolution of the species in biological systems and they have been applied to optimization problems with significant success. However, these algorithms have stochastic nature since the search involves probabilistic rules and the setting of several parameters is difficult. Thus, it is crucial to investigate how the parameters influence the performance of the algorithms for distinct classes of optimization problems. Receiver Operating Characteristic (ROC) analysis, over the years, has become a powerful tool to measure diagnostic performance and, in this work, was used to assist the algorithm parameters setting. The study is conducted over a significant number of difficult optimization problems and a single parameter of an evolutionary algorithm. The ROC analysis seems to be helpful to identify parameters values that allow a satisfactory performance of the algorithm for different classes of problems.

Keywords: Evolutionary Computation, ROC Analysis, Tuning Parameters.

1 Introduction

Several evolutionary approaches have been applied to optimization problems with success, namely Genetic Algorithms [1] and Evolution Strategies [2]. However, Evolutionary Algorithms (EAs) have multiple parameters that are difficult to set since they are problem dependent. On the other hand, in general, interaction between parameters exists and must be taken into account. The usual approach is to perform some previous experimentation (often based on trial and error approaches) in order to tune the algorithm parameters. This task is difficult, inefficient and requires some expertise in order to make EAs work conveniently. Thus, a more systematic and efficient approach that could reduce or avoid this effort will be useful. For this purpose, in this work, ROC analysis was used to assist the tuning of algorithms parameters. ROC analysis can be traced to the theory of statistical decision making, related to applications in signal detection and psychology [5, 6] and, over the years, has become a powerful tool to measure diagnostic performance in medicine [7, 8]. This tool can be used to study the influence of the algorithm parameters on its performance in terms of convergence.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 217–222, 2009.
springerlink.com

In section 2, a short introduction to optimization using evolutionary algorithms is presented. Section 3 describes the ROC Analysis principles. In section 4, the results of the application of ROC analysis to tune an algorithm in order to solve several test problems are presented and discussed. Finally, some conclusions and future work are addressed in section 5.

2 Optimization with Evolutionary Algorithms

Evolutionary algorithms (EAs) mimic the natural evolution of the species in biological systems and they can be used as robust global optimization tools [3]. Moreover, they do not impose any continuity or convexity property of the problem being solved and require only information regarding the objective function. In this work, a particular Evolution Strategy (ES) [2] is considered to solve nonlinear optimization problems without constraints. The goal is to find the global optimum of problems that can be formulated, mathematically, as follows:

$$\min f(x) \text{ with } x \in \mathrm{R}^n \text{ subject to } \alpha \leq x \leq \beta \tag{1}$$

where x is the vector of n real decision variables, $f(x)$ is the objective function to minimize, α and β are the vectors of the lower and upper bounds of the decision variables. The evolutionary approach here considered was the Parameter-less Evolution Strategy (PLES) [4]. In this algorithm an effort was made in order to avoid the difficult task of setting initial values for parameters. However, the parent population size is one of the parameters that must be fixed. The influence of this parameter on convergence will be studied in this work using ROC analysis.

3 ROC Analysis

ROC analysis can be used to compare the performance of diagnostic systems. For a given diagnostic system, the compromise between the False Positive Rate (FPR) and the True Positive Rate (TPR) can be graphically presented through a ROC curve. Thus, each point on the curve defines a compromise, since an increase in TPR is obtained at the cost of an increase on the FPR. The usual measures for testing ROC curves are based on the comparison of the areas under the ROC curves (AUC) that is a global measure of accuracy across all possible threshold values (c):

$$AUC = \int_0^1 ROC(t)dt = \int_{-\infty}^{+\infty} TPR(c)dFPR(c) \tag{2}$$

When comparing two diagnostic systems based on the AUC index, the ROC curve with the greatest value of the index corresponds to the system which presents a better performance, i.e., a greater discriminant power (if the two ROC curves do not cross each other[8]). The graphical representation permits the visualization of the regions of sensitivity and specificity where one curve

is superior to another. This can help to select the off point of the diagnostic systems.

In this work, ROC analysis is used to compare the performance of algorithms when their search parameters are changed. Different combinations of values of parameters can be fixed and they can determine the success of an algorithm to solve a given problem. The best values of the parameters are those that allow a good discrimination between convergence and nonconvergence. Moreover, the corresponding ROC curve allows to identify the tradeoff between TPR and FPR of the algorithm to the parameters settings. Thus, ROC curve is useful to assist the selection of the values of the parameters (the cutoff point) that lead to better algorithm performance. Moreover, AUC is an estimative of the influence of the parameters in the performance (the accuracy) of an algorithm when a particular optimization problem is solved.

4 Results

In this section the ROC analysis for setting parent population size of PLES is presented. Eight test problems, collected by Suganthan (available at the web site address: http://www.ntu.edu.sg/home/EPNSugan), were considered (Table 1). For illustration purposes of the ROC analysis, in this paper only the results for the problems considering 2 decision variables ($n = 2$) are included. The initial populations of the algorithm were uniformly generated at random within the search space.

Table 1. Test Problems

Unimodal	P1: Shifted Sphere Function
	P2: Shifted Schwefel's Problem 1.2
	P4: Shifted Schwefel's Problem 1.2 with Noise in Fitness
Multimodal	P6: Shifted Rosenbrock's Function
	P8: Shifted Rotated Ackley's Function (Global Optimum on Bounds)
	P9: Shifted Rastrigin's Function
	P10: Shifted Rotated Rastrigin's Function
	P11: Shifted Rotated Weierstrass Function

In order to study the influence of parent population size (μ) in the algorithm performance, each problem was solved 25 times considering the following values of the parameter $\mu = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20$. The number of successes (convergences) obtained for each value of μ in each problem was recorded. The convergence criterion was to consider that search was successful when the error ($|f(x) - f(x^*)|$) becomes inferior to 10^{-6} (unimodal problems) or 10^{-2} (multimodal problems), before the maximum number of function evaluations of 10000 was reached.

It should be noted that lower values of μ are more efficient in terms of computational resources requirements. So, it is also desirable to choose the minimum

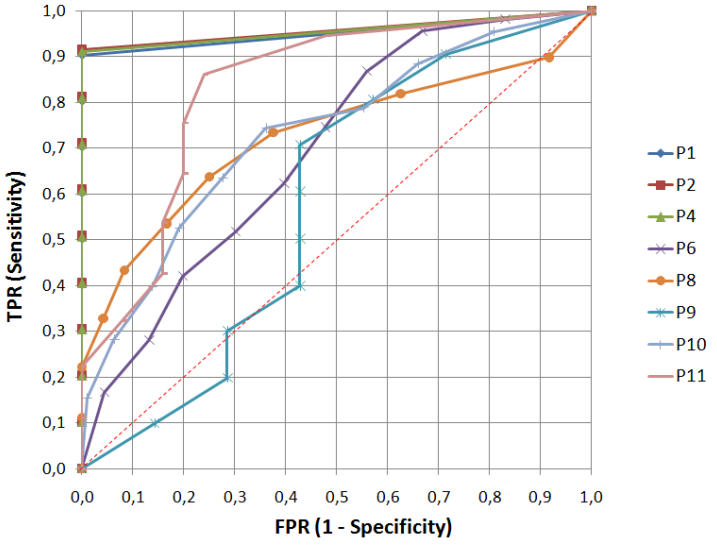


Fig. 1. ROC curves for test problems

value of μ for which the algorithm exhibits a good performance. This statement implies to maximize, simultaneously, sensitivity (TPR) and specificity (1-FPR). TPR can be interpreted as the probability that a convergent example is classified as convergent. Conversely, FPR is the probability that a convergent example is classified as nonconvergent.

The ROC curves obtained for each problem are plotted in Figure 1. Each point on the ROC curve corresponds to a cutoff point, i.e., a particular value of μ. The tradeoff between TPR and FPR is visible and can be considered for the selection of proper values for μ. Moreover, AUC measure can be interpreted as the probability that, when randomly one successful example (convergence occurred) and one unsuccessful example (nonconvergence occurred) are picked, the convergence is correctly predicted by the μ value considered. The most adequate value of μ is the one that discriminates better between convergence and nonconvergence.

Table 2 shows the AUC for all the ROC curves and some hypothetical choices of cutoff points imposing different bounds on the values of FPR. All values of AUC were computed by the trapezoidal rule. It can be observed greater values of AUC for unimodal problems when compared with multimodal problems. So, in these problems, μ has a greater discriminant power. For the multimodal problems, the discriminant power is inferior. The worst AUC was obtained for P9 problem. In this problem the discriminant power between convergence and nonconvergence given by the μ parameter is very low. Thus, in this case, the choice of an initial value of μ seems to have low influence on convergence. The analysis of ROC curves allows to select suitable μ values. Some hypothetical cutoff points

Table 2. Areas Under ROC Curves (AUC) and hypothetical cutoff points

		Cutoff points		
	AUC	FPR< 0.1	FPR< 0.3	FPR< 0.5
P1	0.951	$\mu = 4$ (TPR=0.904)	$\mu = 4$ (TPR=0.904)	$\mu = 4$ (TPR=0.904)
P2	0.957	$\mu = 4$ (TPR=0.915)	$\mu = 4$ (TPR=0.915)	$\mu = 4$ (TPR=0.915)
P4	0.955	$\mu = 4$ (TPR=0.911)	$\mu = 4$ (TPR=0.911)	$\mu = 4$ (TPR=0.911)
P6	0.695	$\mu = 20$ (TPR=0.167)	$\mu = 16$ (TPR=0.421)	$\mu = 10$ (TPR=0.746)
P8	0.726	$\mu = 14$ (TPR=0.434)	$\mu = 10$ (TPR=0.637)	$\mu = 8$ (TPR=0.735)
P9	0.581	$\mu > 20$ (TPR≈ 0)	$\mu = 16$ (TPR=0.300)	$\mu = 8$ (TPR=0.708)
P10	0.730	$\mu = 18$ (TPR=0.282)	$\mu = 12$ (TPR=0.635)	$\mu = 10$ (TPR=0.744)
P11	0.831	$\mu = 16$ (TPR=0.324)	$\mu = 6$ (TPR=0.862)	$\mu = 4$ (TPR=0.947)

are presented in Table 2 representing different compromises between TPR and FPR. From the optimization perspective, the cutoff points presented in the last column of the table seem to be the most adequate. It should be stressed that all these cutoff points present acceptable values of TPR. Taking into account these results, if a FPR threshold not superior to 0.3 is imposed, then the μ value that seems to be more robust for solving a wide number of problems, would be 10 individuals (representing an acceptable compromise as an initial default value for the parameter).

5 Conclusions

A methodology based on ROC analysis that assists the tuning of parameters on evolutionary algorithms was presented. ROC analysis seems to be helpful to identify parameters values that allow a satisfactory performance of algorithms for different classes of problems. This is a relevant issue since it can reduce or avoid the effort of setting parameters based on a trial and error approach. This methodology can also provide some guidelines to select the suitable values of parameters for solving different classes of problems.

Future work will include the study and improvement of this methodology in order to deal with more than one parameter of the algorithm; application to a wide set of optimization problems with distinct features; and the statistical comparison of ROC curves.

References

1. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading (1989)
2. Schwefel, H.-P.: Evolution and Optimum Seeking. Wiley, New York (1995)
3. Costa, L., Oliveira, P.: Evolutionary algorithms approach to the solution of mixed integer non-linear programming problems. Computers Chem. Engng. 25, 257–266 (2001)

4. Costa, L.: A new parameter-less evolution strategy for solving unconstrained global optimization problems. Wseas Transactions on Mathematics 11(5), 1247–1254 (2006)
5. Swets, J.A., Pickett, R.M.: Evaluation of Diagnostic Systems Methods from Signal Detection Theory. Academic Press, London (1982)
6. Swets, J.A.: Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. LEA, New Jersey (1996)
7. Hanley, J.A., McNeil, B.J.: The Meaning and Use of the Area under a Receiver Operating Characteristic ROC Curve. Radiology 143, 29–36 (1985)
8. Metz, C.E.: Statistical Analysis of ROC Data in Evaluating Diagnostic Performance. Herbert D, Myers R, editors. Applications in the Health Sciences. American Institute of Physics 13, 365–384 (1986)

Speeding-Up ACO Implementation by Decreasing the Number of Heuristic Function Evaluations in Feature Selection Problem

Yudel Gómez[1], Rafael Bello[1], Ann Nowé[2], and Frank Bosmans[2]

[1] Department of Computer Science, Universidad Central de Las Villas, Cuba
{ygomezd,rbellop}@uclv.edu.cu
http://www.cei.uclv.edu.cu/
[2] Comp Lab, Department of Computer Science, Vrije Universiteit Brussel, Belgium
{ann.nowe,fpbosmans}@vub.ac.be
http://como.vub.ac.be/

Abstract. In this paper we study a model to feature selection based on Ant Colony Optimization and Rough Set Theory. The algorithm looks for reducts by using ACO as search method and RST offers the heuristic function to measure the quality of one feature subset. Major results of using this approach are shown and others are referenced. Recently, runtime analyses of Ant Colony Optimization algorithms have been studied also. However the efforts are limited to specific classes of problems or simplified algorithm's versions, in particular studying a specific part of the algorithms like the pheromone influence. From another point of view, this paper presents results of applying an improved ACO implementation which focuses on decreasing the number of heuristic function evaluations needed.

1 Introduction

The supervised learning methods applying evolutionary algorithms to generate knowledge model are extremely costly in time and space. Basically, this high computational cost is due to the evaluation process that needs to go through the whole datasets. Often, this process carries out some redundant operations which can be avoided [1].

A meta-heuristic algorithm on the base of ants' behavior was developed in early 1990s by Dorigo, Maniezzo, and Colorni [2]. They called it Ant Colony Optimization [3] because it was motivated by ants' social behavior. Ants are capable of finding the shortest path from food source to their nest or vice versa by smelling pheromones which are chemical substances they leave on the ground while walking. Each ant probabilistically prefers to follow a direction rich in pheromone. This behavior of real ants can be used to explain how they can find a shortest path.

With meta-heuristic ACO, a simulated colony of individual ants moves through the problem space by applying local stochastic rule policy. As they move through the problem space, the ants incrementally build solutions to the problem, evaluating the quality of their partial-step solution, and depositing information on their path in the form of a pheromone message that directs future ants toward iteratively better solutions.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 223–232, 2009.
springerlink.com

The aim of this research is to avoid repetitive evaluation of the heuristic function during the process of choosing next node to incorporate it to the ant's tabu list. In feature selection problem (FSP) it happens repeatedly, not just for each node but for subsets. Experiments show improving is appreciated, decreasing the computing time.

The feature selection is useful in different computational task, for instance, in machine learning processes. Too many irrelevant features increase the complexity of learning process and decrease the accuracy of induced knowledge. Feature selection methods search through the subsets of features and try to find the best subset among the competing 2^N-1 candidate subsets according to some evaluation measure, where N denotes the total number of features.

Bearing in mind FSP as a search space problem, we can consider each state representing a subset of features. All feature selection methods contain two important components: Search algorithm to search through the feature space and an Evaluation function used to evaluate a candidate feature subset. Search strategies are important because the feature selection process may be time consuming and an exhaustive search for the "optimal" subset is impractical for even moderate sized problems [4]. Examples of search strategies are heuristic search [5], probabilistic methods and hybrid algorithms. Another important component is the evaluation function that provides a measure of the quality for the feature subsets.

The approach presented in this paper is based on Rough Sets Theory (RST) and Ant Colony Optimization (ACO). The first is used to build an evaluation function and the second to implement the search method. Methods which combine ACO and Rough Set Theory to find reducts (defined in section 2) with promising results were proposed in [6], [7], [8]. They are based on the reduct concept from RST. Also, the approach includes a measure based on rough sets used to build the evaluation function.

The outline of the paper is as follow. Next section is dedicated to introduce ACO algorithms, Rough Set Theory and generalities about the hybrid model ACO-RST for the feature selection problem are given. Section 3 presents the influence of heavy heuristic functions in ACO's runtime and a modest strategy to avoid repetitive invoking it. Finally, in section 4 we show some experimental results explaining the benefit obtained when applied to feature selection problem.

2 About ACO and RST

2.1 Ant Colony Optimization

Ant colony optimization (ACO) is a population-based metaheuristic that can be used to find approximate solutions to difficult optimization problems [9].

In ACO, at each cycle, a number of artificial ants sequentially construct solutions in a randomized and greedy way. Each ant chooses the next element to be incorporated into its current partial solution on the basis of some heuristic evaluation and the amount of pheromone associated with that element. The former provides the value of any specific candidate solution; the latter represents the memory of the system, and is related to the presence of that element in previously constructed solutions (in the same way as the strength of a pheromone trail is related to how many ants previously chose

to follow that path). Randomization is used to allow the construction of a variety of different solutions. Basically, a probability distribution is defined over all elements which may be incorporated into the current partial solution, with a bias in favor of the best elements. In particular, an element with a good heuristic evaluation and a high level of pheromone is more likely to be selected according to expression (1).

$$p_{ij}^{k} = \begin{cases} 0 & if \quad j \notin N_i^k \quad (\text{neighborhood of ant k}) \\ \dfrac{(\tau_{ij})^{\alpha} \cdot (\eta_{ij})^{\beta}}{\sum_{l \in N_i^k} (\tau_{il})^{\alpha} \cdot (\eta_{il})^{\beta}} & if \quad j \in N_i^k \end{cases} \tag{1}$$

Each time an element is selected by an ant, its pheromone level is updated by first removing a fraction of it, to mimic pheromone evaporation, and then by adding some new pheromone. When all ants have constructed a complete solution, the procedure is restarted with the updated pheromone levels. This is repeated for a fixed number of cycles or until search stagnation occurs.

Some direct successor algorithms of Ant Systems are: Elitist AS, Rank-based AS and *MAX-MIN* AS. Another ACO algorithm is Ant Colony System (ACS). ACS uses an extra pseudorandom proportional rule to select the next node j from node i, see expression (2). And change slightly the form to update the pheromone.

$$j = \begin{cases} \arg\max_{l \in N_i^k} \left\{ \tau_{ij} \cdot (\eta_{il})^{\beta} \right\} & if \quad q \le q_0 \\ random\ selection\ according\ to\ (1) & otherwise \end{cases} \tag{2}$$

2.2 Rough Sets Theory

Rough Sets Theory (RST) was proposed by Z. Pawlak [10]. The rough set philosophy is founded on the assumption that some information is associated with every object of the universe of discourse. Rough set data analysis is one of the main techniques arising from RST; it provides a technique for gaining insights into the data properties. The rough set model has several advantages for data analysis. It is based on the original data only and does not need any external information; no assumptions about data are made; it is suitable for analyzing both quantitative and qualitative features, and the results of rough set model are easy to interpret [11].

In RST a training set can be represented by a table where each row represents objects and each column represents an attribute. This table is called an Information System; more formally, it is a pair S= (U, A), where U is a non-empty finite set of objects called the Universe and A is a non-empty finite set of attributes. A Decision System is a pair DS=(U, A$\cup${d}), where d$\notin$ A is the decision feature. The basic concepts of RST are the lower and upper approximations of a subset X$\subseteq$U [12]. These were originally introduced with reference to an indiscernibility relation IND(B), where objects x and y belong to IND(B) if and only if x and y are indiscernible from each other by features in B.

Let B$\subseteq$A and X$\subseteq$U, it can be proved that B defines an equivalence relation. The set X can be approximated using only the information contained in B by constructing the

B-lower and B-upper approximations of X, denoted by B_*X and B^*X respectively, where $B_*X=\{x : [x]_B \subseteq X\}$ and $B^*X=\{x : [x]_B \cap X \neq \phi\}$, and $[x]_B$ denotes the class of x according to B-indiscernible relation. The objects in B_*X are guaranteed to be members of X, while the objects in B^*X are possibly members of X. If B^*X-B_*X is not empty, then X is a rough set.

RST offers several measures about a Decision System. Among them is the quality of the approximation of classification (expression 1). It expresses the percentage of objects which are correctly classified into the given classes $Y=\{Y_1, ..., Y_n\}$ employing only the set of features in B.

$$\gamma_B(Y) = \frac{\sum_{i=1}^{n} |B_* Y_i|}{|U|} \tag{3}$$

An important issue in the RST is feature reduction based on the reduct concept. A reduct is a minimal set of features that preserves the partitioning of universe and hence the ability to perform classifications [12]. The subset B is a reduct if IND(A)=IND(B); that is, $\gamma_A(Y)=\gamma_B(Y)$. The practical use is limited because of the computational complexity of calculating reducts. The problem of finding a globally minimal reduct for a given information system is NP-hard. In [13], Bell shows that the computational cost of finding a reduct in the information system that is limited by l^2m^2, where l is the length of X and m is the amount of objects in the universe of the information system; while the complexity in time of finding all the reducts of X is $O(2^lJ)$, where l is the amount of attributes and J is the computational cost required to find a reduct. For those reasons, methods for calculating reducts have been developed using heuristic methods [14],[6],[15].

2.3 The Hybrid Model Used for Feature Selection (ACO-RST-FS)

Feature selection problem is an example of a difficult discrete problem which can be represented as a graph problem; for that reason, the ACO approach is interesting to solve it. There are at least two interesting ACO approach to solve it presented by Jensen and Chen [6] and Bello [7], [8]. In this research we follow the hybrid model presented by Bello, the search method is based on the ant approach and the evaluation heuristic function is provided by RST using the quality of the approximation of classification measure. Experimental results have showed this hybrid approach allows to obtain shortest reducts, and the high computational cost has encouraged finding strategies to decrease the runtime.

Model Generalities

The graph includes a node for each feature. Each node (attribute a_i) has associated a pheromone value τ_i and all nodes are connected by bidirectional links. Each ant builds a subset of features step by step denoted by b_k, where k refers to the ant k^{th}. In the initial step of each cycle ants are associated to nodes in the graph. In each next step, ants select another feature to include in their subsets (used also as a tabu list). The ACO algorithm uses the quality of the approximation of the classification as heuristic

function (η) to evaluate a subset B, (η(B)= γ_B(Y)). It looks for subsets B such that γ_B(Y)= γ_A(Y). The resulting expression for the random proportional rule stays as:

$$p_k\left(B^k, a_j\right)=\begin{cases} 0 & if \quad a_j \in B^k \\ \dfrac{\left[\tau_j\right]^\alpha \cdot \left[\gamma_{B^k \cup \{a_j\}}(Y)\right]^\beta}{\sum_{a_j \in A-B^k} \left[\tau_j\right]^\alpha \cdot \left[\gamma_{B^k \cup \{a_j\}}(Y)\right]^\beta} & if \quad a_j \in A-B^k \end{cases} \tag{4}$$

Similar changes in the heuristic function will be applied to expression (2).

3 Improving ACO Runtime

Recently, runtime analyses of Ant Colony Optimization algorithms have been also studied, but the efforts are limited to specific classes of problems or simplified algorithm's versions, in particular studying a piece of the algorithms like pheromone influence [16], [17]. However, an element that introduces high computational cost is the calculation of the heuristic function. As suggested in the literature [18], [19] we focus on the runtime analysis by the number of necessary function evaluation too.

Ants select the next node according to expression (1) or (2) (depending on AS or ACS), calling heuristic function to calculate the quality of the nodes. For single heuristics functions with low computational complexity it is not essential to study ACO runtime including heuristic analysis. But, for heuristic functions with high computational complexity it has high influence in the model complexity and runtime. We can classify the problem which can be solved with ACO algorithms taking into account the way of heuristic evaluation: one takes just information about the next node, such as the distance to the next city in the TSP problem, one node each time to evaluate the heuristic; the other takes no just the next node to include in the solution but also the tabu list remaining in the ant, such as Set Covering Problem. In the last case, the heuristic information depend on the next node and the subset of nodes visited by the ant; if we could have information about those subsets of nodes we could try to save reiterative evaluations and save a lot of computing time.

As solution, we propose to use a data structure which stores all evaluated subset and the quality (heuristic value) calculated; then every called to the function first look for the subset in the data structure. If the subset is found the associated value is returned else the heuristic function is evaluated and both the subset and the value are stored in the data structure for future use.

4 Experimental Results

In earlier work [6], R. Jensen and Q. Shen have proposed a method which combine Ant model and Rough sets (AntRSAR) to find reducts, they showed promising results when comparing the results of this method with other methods to build reducts such as RSAR (using Quickreduct) and GenRSAR (genetic algorithm based). In Table 2, we present a comparison between ACS-RST-FS, AntRSAR, RSAR and GenRSAR

Table 1. Comparison with other algorithms using data base from UCI Repository

Data base	objects	features	RSAR	GenRSAR	AntRSAR	ACS-RST-FS
Heart	294	13	7	6	6	6
Exactly	1000	13	9	6	6	6
Exactly 2	1000	13	13	10	10	10
LED	2000	24	12	6	5	5
Lung Cancer	32	56	4	6	4	4

algorithms (data for the last three from [6]) with respect to the length of minimal reduct found by algorithms. These results show that the algorithms based on ant models obtain the best results. For ACS-RST-FS, 21 cycles were enough to obtain these results while the results reported back in [6] were obtained in 250 cycles.

In [7] and [8] authors show a comparison between ACS-RST-FS and other algorithms analyzing parameters like quantity of reducts, average length and how many reducts are found by the algorithm with minimal length. Also a study about of parameters of the ACO metaheuristic for the FSP is presented.

In this research we incorporated the strategy described in topic 3 in the model ACS-RST-FS using a single list as data structure. To evaluate the effect of introducing this new strategy we run the algorithm and count, for every cycle, the amount of times the heuristic equation was evaluated, given by the expression (3) (column 2, Table 2.), how many times we found in the list the heuristic values that are already calculated (column 3, Table 2.); the addition of these two quantities corresponds to the total of evaluated candidate subsets (column 4, Table 2.). This experiment was carried out with different datasets from the repository UCI [20]. Table 2. shows the results obtained for Dermatology dataset during 15 cycles with 15 ants. Also, another experiment was carried out with Breathcancer, Ledjen, Lung-cancer and Heart-disease for15 and 150 cycles having, in all cases, similar behaviour. These results are not given here.

Fig. 1. graphics depicts the information of table 2. The column ii of Table 2., plotted as a dotted line, represents the number of times the heuristic function was evaluated by the expression (3). The dashed plot, column iii of Table 2. represents the number of subsets for which quality was already evaluated and stored in the data structure. The solid plot, column iv of Table 2. , presents the totality of sets to be evaluated.

The total number of candidates that need to be evaluated per cycle shows an oscillatory behaviour, due to the stochastic character of the ACO metaheuristic.

As can be observed from graph, in every new cycle the number of evaluations of expression (3) tends to decrease. The dotted line decreases with the number of cycles. Note that this line corresponds to the difference between the dashed line and the solid line. This happens due to the growing of the number of subsets found within data the structure during the execution of the algorithm; considering that the expansion in this data structure and the complexity to retrieve the quality of a subset within it will never be more expensive than evaluating the expression (3).

Table 2. Number of Subsets evaluated in each cycle

Cycle (i)	Calculated Expression (3) (ii)	Stored in the data structure (iii)	Total candidate subsets (iv)
1	3426	11673	15099
2	2664	7940	10604
3	1993	6498	8491
4	2238	13125	15363
5	1385	10221	11606
6	1432	14936	16368
7	1566	11806	13372
8	1055	9985	11040
9	752	11541	12293
10	950	9472	10422
11	592	12836	13428
12	729	11687	12416
13	502	17314	17816
14	709	11234	11943
15	349	9079	9428

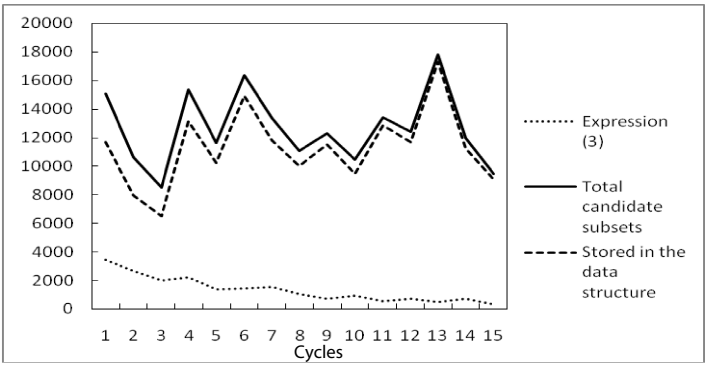


Fig. 1. Subset evaluations

The fact that progressively more subsets are evaluated out of the information found in the list instead of evaluating the heuristic function given by the expression (3) makes the computational cost in time decrease.

The complexity analysis supports what has been stated before.

Let be:

I number of iterations of the algorithm, *n* quantity of ants, *l* number of original features, *m* amount of objects in the dataset, *c* number of classes in the decision system and *s* the amount of elements stored in the list

Time complexity for the ACO-FS [21] $O(I \cdot n \cdot l)$

Complexity to evaluate expression (3) $O(c \cdot l^2 \cdot m^2)$

Complexity to find a set in the list $O(l \cdot s)$

The number of elements stored in the list increases dynamically according to the stochastic character of ACO algorithms. In the first cycles the growth will be quick but then it will grow slowly because many subsets will already be in the list. In every cycle the highest number of elements in the list is of the order $n*l^2$, so $c*m^2$ has an order greater than $l*n$, in consequence the complexity to evaluate expression (3) has an order greater than search a set in the list.

For those reasons, it is likely that the expected runtime until reaching the optimal solution is even of a more favorable order using the proposed strategy. Another interesting aspect to take into account is the comparative analysis of the search space explored by ACO-RST-FS algorithm versus an exhaustive method.

Table 3. Search space: ACO-RST-FS algorithm vs exhaustive method

Dataset	Features	Reducts		ACS-RST-FS		Exhaustive search
		Length Average	Shortest subset	Calculated expression (3)	Stored in the data structure	Search Space
(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
Heart	13	10.2	7	943	33020	10057645
Led	24	7.7	5	2802	622160	5368224
Lung	56	5.7	4	21310	731118	8984416
Breast Cancer	9	5.4	4	251	7442	3609
Dermatology	34	20.6	10	26291	646980	4.956E+14

Table 3. shows calculations accomplished for datasets from UCI Repository. Columns (iii) and (iv) illustrate the number of features in the reducts found by the algorithm ACO-RST-FS, column (iii) length average, column (iv) shortest length gotten. Column (v) y (vi) represent the number of subset evaluated by ACO-RST-FS algorithm to obtain the reducts and column (vii) represents the quantity of subsets that should evaluate an exhaustive method, that is the quantity of nodes of the search space that should visit. Column (vii) is calculated by expression (5)

$$Number\_of\_Nodes = \sum_{l=1}^{MIN} Subsets(length = l) \tag{5}$$

where:

Subsets(length=l) is the number of subsets with l features and MIN is the number of features in the shortest subset.

In general for datasets of medium and big size the search space of an exhaustive method until reaching an optimal solution will be much bigger than the search space explored by the heuristic method.

5 Conclusions

We have presented the possibilities of applying a hybrid model: ACO and Rough Sets to feature selection. As important issue this research includes an improved ACO

implementation which focuses on decreasing the number of heuristic function evaluations needed to speed-up the runtime. Principal advantage of this model is to find several subsets of features, as solutions, at the same time; different of other approaches. The comparison of the hybrid model with other heuristic methods to calculate reducts in RST shows that ant model based algorithms yield good results.

According to the study of the behavior of this ACO implementation we can establish that the exploration of the search space is a combination of best first and depth first search like a kind of beam search. As well we can orient future work to investigate how ants can explore the same part of search space or the ants can be spread over the search space; depending on if there is only one minimal reduct or the reducts are strongly overlapped, or there are several reducts which do not overlap.

References

1. Giráldez, R., Díaz-Díaz, N., Nepomuceno, I., Aguilar-Ruiz, J.S.: An Approach to Reduce the cost of Evaluation in Evolutionary Learning. In: Cabestany, J., Gonzalez Prieto, A., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 804–811. Springer, Heidelberg (2005)
2. Dorigo, M., DiCaro, G., Gambardella, L.M.: Ant colonies for discrete optimization. Artificial Life 5, 137–172 (1999)
3. Dorigo, M., Stutzle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
4. Zhang, H., Sun, G.: Feature selection using tabu search method. Pattern Recognition Letters 35, 710–711 (2002)
5. Silver, E.: An overview of heuristic solution methods. Journal of the Operational Research Society 55, 936–956 (2004)
6. Jensen, R., Shen, Q.: Finding Rough Set Reducts with Ant Colony Optimization. In: UK Workshop on Computational Intelligence, 15–22 (2003)
7. Bello, R., Nowé, A.: A Model based on Ant Colony System and Rough Set Theory to Feature Selection. In: Genetic and Evolutionary Computation Conference (GECCO 2005), pp. 275–276 (2005)
8. Bello, R., Nowé, A.: Using Ant Colony System meta-heuristic and Rough Set Theory to Fea-ture Selection. In: The 6th Metaheuristics International Conference (MIC 2005), Vienna, Austria (2005)
9. Dorigo, M.: Scolarpedia, vol. 2 (2007)
10. Pawlak, Z.: Rough sets. International Journal of Information & Computer Sciences 11, 341–356 (1982)
11. Tay, F.E.S.L.: Economic and financial prediction using rough set model. European Journal of Operational Research 141, 641–659 (2002)
12. Komorowski, J.a.P.Z.: Rough Set: A tutorial. Rough Fuzzy Hybridization: A new trend in decision making, 3–98 (1999)
13. Bell, D., Guan, J.: Computational methods for rough classification and discovery. Journal of ASIS 49, 403–414 (1998)
14. Wroblewski, J.: Genetic algorithms in decomposition and classification problems. In: Polkowski, L., Skowron, A. (eds.) Rough sets in Knowledge Discovery 1: Applications, pp. 472–492. Physica-Verlag
15. Wang, X.: Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters 28, 459–471 (2007)

16. Doerr, B., Neumann, F., Sudholt, D., Witt, C.: On the Runtime Analysis of the 1-ANT ACO Algorithm. In: GECCO 2007 (2007)
17. Neumann, F., Witt, C.: Runtime analysis of a simple Ant Colony Optimization algorithm. In: Asano, T. (ed.) ISAAC 2006. LNCS, vol. 4288, pp. 618–627. Springer, Heidelberg (2006)
18. Gutjahr, W.J.: First steps to the runtime complexity analysis of ant colony optimization. Computers and Operations Research (2007)
19. Droste, S., Jansen, T., Wegener, I.: On the analysis of the (1+1) evolutionary algorithm. Theoretical Computer Science 276, 51–81 (2002)
20. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998), http://www.ics.uci.edu/~mlearn/MLRepository.html
21. Jensen, R., Shen, Q.: Fuzzy-Rough Data Reduction with Ant Colony Optimization. Fuzzy Set and System 149, 5–20 (2005)

Global Sensitivity Analysis of a Biochemical Pathway Model

Maria Rodriguez-Fernandez and Julio R. Banga

Process Engineering Group.
Instituto de Investigaciones Marinas-CSIC,
C/Eduardo Cabello 6, 36208. Vigo spain
mrodriguez@iim.csic.es, julio@iim.csic.es

Summary. Developing suitable dynamic models of biochemical pathways is a key issue in Systems Biology. Parameter identification is therefore a critical aspect where the analysis of the model identifiability plays an important role. The study of model identifiability aims to determine whether the unknown parameters can be uniquely estimated from the available experiments. As the number of parameters and the complexity of the models increase, current methods for testing structural identifiability become inapplicable, so methods mostly based on local sensitivities are frequently used. Although these methods are valid for linear cases, or when the value of the parameters is known, they can be misleading for the general nonlinear case. Parametric global sensitivity analysis is presented here as a robust alternative for this type of analysis, detecting non influential parameters and their interactions. The performance of this methodology is illustrated with a benchmark dynamic model describing a biochemical pathway.

Keywords: Identifiability analysis, global sensitivity analysis, Sobol' indices, derivative based sensitivity measures.

1 Introduction

Dynamic computational models are powerful tools for developing and testing hypotheses about complex biological systems. However, a major challenge with such models, is that they often posses tens or even hundreds of free parameters whose values must be estimated from noisy and often scarce experimental data.

A rough optimization of the agreement between the model predictions and the available data may lead to large parameter uncertainties. Thus, a further analysis of the model structure should be performed prior the parameter estimation.

In this work, deterministic nonlinear dynamic models of biochemical pathways, i.e. those described by deterministic ordinary differential equations (ODEs), differential-algebraic equations (DAEs), or partial differential equations (PDEs), are considered. In the case of ODEs, a popular statement is the so-called state-space formulation:

$$\dot{x}(p,t) = f[x(p,t), u(t), p],\ x(0) = x_0, \tag{1}$$

$$y(p,t) = g[x(p,t), u(p,t), p] \tag{2}$$

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 233–242, 2009.
springerlink.com

where x is the vector of N_x state variables and p the vector of n model parameters. Note that f specify the model, u specifies the vector of inputs (i.e. for a particular experiment) and y the vector of N_y measured states. An experiment is specified by the initial conditions $x(0)$, the inputs u chosen among a set of possible inputs U and the observations y. Note that the inputs can be time-dependent.

Sensitivity analysis quantifies the dependence of the system behavior on the parameters that affect the process dynamics. If a small change in a parameter results in relatively large changes in the state variables, the model answer is said to be sensitive to this parameter. This means that this type of parameters have to be determined very precisely to achieve an accurate model. On the contrary, parameters to which the process dynamics exhibit a small sensitivity do not need to be measured rigourously.

There are several sensitivity analysis methods than can be classified as: screening methods, local sensitivity analysis methods and global sensitivity analysis methods [5]. This distinction is somewhat arbitrary, since screening test can also be viewed as either local or global. Further, the first class is characterized with respect to its use (screening), while the other two are characterized with respect to how they treat factors.

In the context of numerical modeling, sensitivity indices play an important role in uncertainty analysis, parameter estimation and optimisation, experimental data analysis, and model discrimination. The results of a sensitivity analysis can be used to:

- validate a model
- warn of strange or unrealistic model behavior
- suggest new experiments or guide future data collection efforts
- point out important assumptions of the model
- suggest the accuracy to which the parameters must be estimated
- guide the formulation of the structure of the model
- adjust numerical values for the parameters

In the present contribution, a new parametric global sensitivity analysis method is extended to be able to handle DAEs systems and applied to a sound benchmark problem. The results are compared with the local sensitivities and the Sobol' sensitivity indices.

2 Global Sensitivity Analysis

Local sensitivity indices are computed at some nominal values considered for the parameters and the behavior of the response function is described only locally in the input space. Moreover, preliminary experiments and model calibration tests should be carried out in order to obtain a first guess for the parameter values and an iterative scheme involving both steps is required in order to study the model sensitivity. In addition, these methods are linear thus they are not sufficient for dealing with complex models, especially those in which there are nonlinear interactions between parameters.

In contrast, global sensitivity analysis (GSA) methods evaluate the effect of a parameter while all other parameters are varied simultaneously, thus accounting for interactions between parameters without depending on the stipulation of a nominal point (they explore the entire range of each parameter).

The most widely used methods in GSA are, FAST and extended FAST, the Morris method and its adaptations and the Sobol' method considered as one of the more powerful despite its high computational cost.

2.1 Sobol' Global Sensitivity Indices

The method of global sensitivity indices developed by Sobol' is the most established among the variance-based methods. The method is based on the ANOVA decomposition of the variance of the model output. A detailed description of the method can be found in [7].

Sobol' [7] define two type of indices: S_{p_a} that accounts only for the effect of the parameters of the subset a and $S^T_{p_a}$ that also accounts for the interactions of the parameters of the subset a with the rest of the parameters. He found a elegant way of computing these indices directly from the model $f(p)$:

$$S_{p_a} = \frac{\int_{H^n} f(p_a, p_b) f(p_a, p'_b) dp_a dp_b dp'_b - f_0^2(p)}{\int_{H^n} f^2(p) dp - f_0^2(p)} \tag{3}$$

$$S^T_{p_a} = \frac{1}{2} \frac{\int_{H^n} \left[f(p_a, p_b) f(p'_a, p_b) \right]^2 dp_a dp'_a dp_b}{\int_{H^n} f^2(p) dp - f_0^2(p)} \tag{4}$$

Using these indices a parameter ranking can be established.

2.2 Derivative Based Global Sensitivity Measures

Kucherenko and co-workers [2] presented the derivative based global sensitivity measures (DGSM) based on averaging local derivatives using Quasi Monte Carlo sampling methods. They applied this technique to a set of explicit functions showing that it is much more accurate than the Morris method. Moreover, they demonstrated that there is a link between these measures and the Sobol' sensitivity indices.

In this work, these measures are extended in order to be able to handle differential-algebraic equations (DAEs). The details of this methodology are described below.

Consider a differentiable function $f(p)$, where $p = \{p_i\}$ is a vector of parameters defined in the unit hypercube H^n $(0 \leq p_i \leq 1, i = 1, ..., n)$. Local sensitivity measures are based on partial derivatives

$$E_i\left(p^*\right) = \frac{\partial f}{\partial p_i} \tag{5}$$

Sensitivity measure $E_i\left(p^*\right)$ depends on a nominal point and it changes with a change of p^*. This deficiency can be overcome by averaging $E_i\left(p^*\right)$ over the parameter space H^n. Such a measure can be defined as:

$$\bar{M}_i = \int_{H^n} E_i dp \tag{6}$$

Another measure, which is the variance of $\bar{M}_i$, is also considered

$$\bar{\Sigma}_i = \left[\int_{H^n} \left(E_i - \bar{M}_i \right)^2 dp \right]^{1/2} \tag{7}$$

$\bar{\Sigma}_i$ can also be presented as

$$\bar{\Sigma}_i^2 = \int_{H^n} E_i^2 dp - \bar{M}_i^2 \tag{8}$$

Combining $\bar{M}_i$ and $\bar{\Sigma}_i$ a new measure $\bar{G}_i$ can be introduced

$$\bar{G}_i = \bar{\Sigma}_i^2 + \bar{M}_i^2 = \int_{H^n} E_i^2 dp \tag{9}$$

Being s a number of parameters in a subset we can define the so called "alternative global sensitivity estimator", $\bar{G}_s$, as

$$\bar{G}_s = \frac{\sum_{i=1}^{s} G_i}{\sum_{i=1}^{n} G_i} \tag{10}$$

Non-monotonic functions have regions of positive and negative values of partial derivatives $E_i(p^*)$, hence due to the effect of averaging values $\bar{M}_i$ can be very small or even zero: i.e. for a symmetrical at a middle point ($p = 0.5$) function $\bar{M}_i = 0$. To avoid such situations measures based on the absolute value of $|E_i(p^*)|$ can be used:

$$\bar{M}_i^* = \int_{H^n} |E_i| \, dp \tag{11}$$

$$\bar{\Sigma}_i^* = \left[\int_{H^n} \left(|E_i| - \bar{M}_i^* \right)^2 dp \right]^{1/2} \tag{12}$$

Similar measures were introduced in [1] within the framework of the Morris method. Using an analogy with variance based global sensitivity measures, the set of measures $\bar{M}_i$, $\bar{\Sigma}_i$ and $\bar{G}_i$ are called derivative based global sensitivity measures (DGSM).

2.3 Computational Algorithms for Calculation of Integrals

Calculation of Sobol' indices and DGSM is based on the evaluation of a series of integrals (those involved in (3)-(4) and (6)-(12) respectively) that can be presented in the following generic form:

$$I[f] = \int_{H^n} f(p) dp \tag{13}$$

It is assumed that function $f(p)$ is integrable in the n-dimensional unit hypercube H^n.

Classical grid methods become inefficient in high-dimensions because of the "curse of dimensionality" (exponential grows of the required integrand evaluations). Monte Carlo methods do not depend on the dimensionality and are effective in high dimensional integrations. However, the efficiency of MC methods is determined by the properties of random numbers. It is known that random number sampling is prone to clustering: for any sampling there are always empty areas as well as regions in which random points are wasted due to clustering. As new points are added randomly, they do not necessarily fill the gaps between already sampled points.

A higher rate of convergence can be obtained by using deterministic uniformly distributed sequences also known as low-discrepancy sequences (LDS) instead of pseudo-random numbers. Methods based on the usage of such sequences are known as Quasi Monte Carlo (QMC) methods.

LDS are specifically designed to place sample points as uniformly as possible. Unlike random numbers, successive LDS points "know" about the position of previously sampled points and "fill" the gaps between them. LDS are also known as quasi random numbers. The QMC algorithm for the evaluation of the integral (13) has a form

$$I_N = \frac{1}{N}\sum_{i=1}^{N} f(q_i) \tag{14}$$

where q_i is a set of LDS points uniformly distributed in a unit hypercube H^n, $q_i = (q_i^1, ..., q_i^n)$.

There are a few well-known and commonly used LDS. Different principles were used for their construction by Holton, Faure, Sobol', Niederreiter and others. Many practical studies have proven that the Sobol' LDS is in many aspects superior to other LDS [6]. For this reason it was used in this work.

For the best known LDS the estimate for the rate of convergence $I_N \to I$ is known to be $O(ln^n N)/N$. This rate of convergence is much faster than that for the MC method, although it depends on the dimensionality n.

Numerical Computation of the Sobol' Indices

From (3), and applying the QMC algorithm for the evaluation of the integrals, the Sobol' indices can be calculated in a straightforward manner according to the formulas:

$$S_{p_a} = \frac{\frac{1}{N}\sum f(p_a, p_b) f(p_a, p'_b) - (\frac{1}{N}\sum f(p))^2}{\frac{1}{N}\sum f^2(p) - (\frac{1}{N}\sum f(p))^2} \tag{15}$$

$$S^T_{p_a} = \frac{1}{2}\frac{\sum [f(p_a, p_b) f(p'_a, p_b)]^2}{\frac{1}{N}\sum f^2(p) - (\frac{1}{N}\sum f(p))^2} \tag{16}$$

Thus, each Quasi Monte Carlo sample point requires three computations of the model $f(p_a, p_b)$, $f(p'_a, p_b)$ and $f(p_a, p'_b)$. For the computation of the Sobol'

indices of an entire set of n parameters, using N sample points, the number of function evaluations is $N_F = N(n+2)$.

Numerical Computation of the DGSM Measures

Evaluation of DGSM measures requires calculation of $E_i(p^*)$. In [2] it is calculated analytically for easy-differentiable functions or numerically using finite difference approximation:

$$E_i(p^*) = \frac{\left[f(p_1^*, ..., p_{i-1}^*, p_i^* + \delta, p_{i+1}^*, ..., p_n^*) - f(p^*)\right]}{\delta} \tag{17}$$

where δ is a small increment.

Proper selection of scalar δ is crucial to maintaining acceptable round-off and truncation error levels. The total number of function evaluation for calculation of a full set of $\bar{M}_i$ and $\bar{\Sigma}_i$ is $N_F = N(n+1)$.

In order to reduce the number of function evaluations and to increase the precision of the sensitivity measures, the use of the iterative approximation to compute the partial derivatives of the DAEs systems is suggested in this work. This approach is based on directional derivatives, in a similar form to that used by Maly and Petzold [3]. The sensitivity equations are solved simultaneously with the original system avoiding the difficult task of selecting a proper δ and reducing the number of function evaluations to $N_F = N$.

2.4 Extension to DAEs Systems

Since dealing with a systems of DAEs, the sensitivity indices of every observed state variable at each measurement time point with respect to each of the parameters, are available. In order to summarize all this information, global sensitivity indices as the average of all the S_i for each parameter are defined:

$$S_i = \frac{1}{N_y}\frac{1}{N_t}\sum_{j=1}^{N_y}\sum_{k=1}^{N_t} S_{ij}(t_k) \tag{18}$$

The same expression is applicable to S_i^T, $\bar{M}_i$, $\bar{\Sigma}_i$ and $\bar{G}_i$.

3 Results for a Benchmark Pathway

3.1 Statement of the Problem

A challenging benchmark problem, involving a biochemical pathway with three enzymatic steps, was considered. The mathematical formulation of the nonlinear dynamic model is:

$$\frac{dG_1}{dt} = \frac{V_1}{1 + \left(\frac{P}{Ki_1}\right)^{ni_1} + \left(\frac{Ka_1}{S}\right)^{na_1}} - k_1\ G_1 \tag{19}$$

$$\frac{dG_2}{dt} = \frac{V_2}{1 + \left(\frac{P}{Ki_2}\right)^{ni_2} + \left(\frac{Ka_2}{M_1}\right)^{na_2}} - k_2\ G_2 \tag{20}$$

$$\frac{dG_3}{dt} = \frac{V_3}{1 + \left(\frac{P}{Ki_3}\right)^{ni_3} + \left(\frac{Ka_3}{M_2}\right)^{na_3}} - k_3\ G_3 \tag{21}$$

$$\frac{dE_1}{dt} = \frac{V_4\ G_1}{K_4 + G_1} - k_4\ E_1 \tag{22}$$

$$\frac{dE_2}{dt} = \frac{V_5\ G_2}{K_5 + G_2} - k_5\ E_2 \tag{23}$$

$$\frac{dE_3}{dt} = \frac{V_6\ G_3}{K_6 + G_3} - k_6\ E_3 \tag{24}$$

$$\frac{dM_1}{dt} = \frac{kcat_1\ E_1\ \left(\frac{1}{Km_1}\right)(S - M_1)}{1 + \frac{S}{Km_1} + \frac{M_1}{Km_2}} - \frac{kcat_2\ E_2\ \left(\frac{1}{Km_3}\right)(M_1 - M_2)}{1 + \frac{M_1}{Km_3} + \frac{M_2}{Km_4}} \tag{25}$$

$$\frac{dM_2}{dt} = \frac{kcat_2\ E_2\ \left(\frac{1}{Km_3}\right)(M_1 - M_2)}{1 + \frac{M_1}{Km_3} + \frac{M_2}{Km_4}} - \frac{kcat_3\ E_3\ \left(\frac{1}{Km_5}\right)(M_2 - P)}{1 + \frac{M_2}{Km_5} + \frac{P}{Km_6}} \tag{26}$$

where M_1, M_2, E_1, E_2, E_3, G_1, G_2 and G_3 represent the concentration of the species involved in the different biochemical reactions.

The identification problem studied in [4] consists of the estimation of 36 kinetic parameters of the nonlinear biochemical dynamic model (8 nonlinear ODEs) which describes the variation of the metabolite concentration with time.

Table 1. Lower and upper bounds for the 36 parameters

	Lower bound	Upper bound
Set I	$10^{-1}p^*$	$10^{1}p^*$
Set II	$10^{-3}p^*$	$10^{3}p^*$

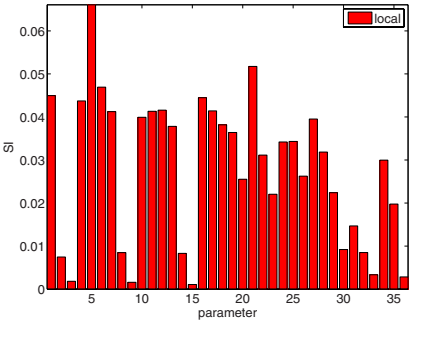


Fig. 1. Local sensitivities at p^*

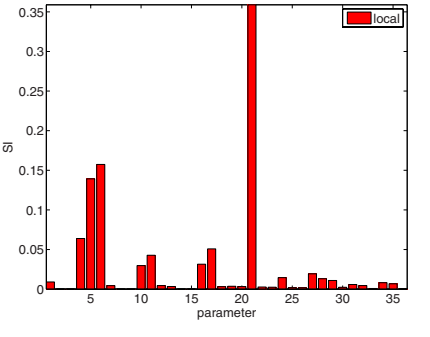


Fig. 2. Local sensitivities at $5p^*$

In order to get a better understanding of the model structure, the local and global sensitivity measures for the entire set of 36 parameters were computed. The local sensitivities were computed at a nominal point p^* and at $5p^*$. Although normalized between 0 and 1, a range for the feasible values of the parameters must be defined to apply the global sensitivity methods. To study the influence of the bounds on the global sensitivity indices, two sets of bounds were considered (see Table 1).

3.2 Results of the Local Sensitivity Analysis

The results of the local sensitivity analysis are shown in Figures 1-2.

With these results a ranking of importance for the parameters was established. For the nominal point, p^*, the ten less influential parameters are p_2, p_3, p_8, p_9, p_{14}, p_{15}, p_{30}, p_{32} p_{33} and p_{36}. They account for the 5 % of information while the ten more influential ones represent the 50 % of the model sensitivity. That indicates that our efforts must be focused on the proper estimation of the most influential parameters whereas the less important can be fixed to their nominal values.

However, when computing the same ranking for a different point, as $5p^*$, the results are completely different. Since the sensitivity analysis is usually performed before the parameter estimation, more robust methods are needed in order to ensure that small changes in the parameters values will not significantly change the results of the study.

3.3 Results of the Global Sensitivity Analysis

The values of the total Sobol' indices and the $\bar{G}_i$ measure were computed using 2^{10} sample points and a ranking of importance was established.

The results for the set of bounds I are shown in Figures 3-4. From the ten less influential parameters, nine are coincident (p_2, p_3, p_8, p_9, p_{14}, p_{15}, p_{25}, p_{33} and p_{36}) showing that $\bar{G}_i$ is a good proxy to S_i^T. Moreover, it can be seen that one third of the parameters accounts for the 50 % of the model sensitivity.

Besides, eight are in agreement with the results of the local sensitivities at the nominal point but no connection can be established with the results for the vector $5p^*$ although this point is included in the considered set of bounds.

As can be seen from the results for the set of bounds II (Figures 5-6), the sensitivity indices obtained using too wide bounds can be misleading.

The results for the Sobol' indices and the DGSM are completely different, both showing an unrealistic situation where only a few parameters account for all the system information. This fact can be due to the lack of convergence of the integration algorithm. Since the bounds are very wide a higher number of sample points should be used increasing exponentially the computational effort.

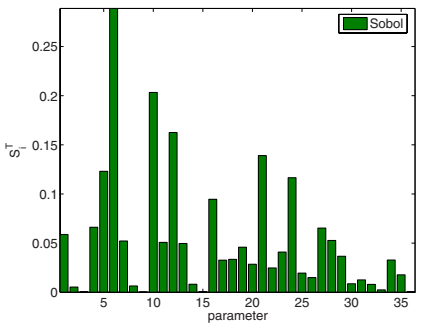


Fig. 3. Sobol', Set of bounds I

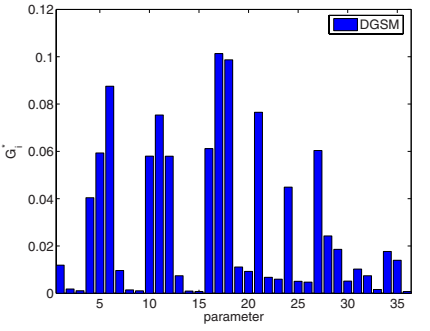


Fig. 4. DGSM, Set of bounds I

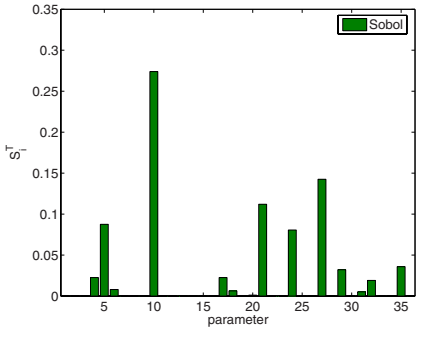


Fig. 5. Sobol', Set of bounds II

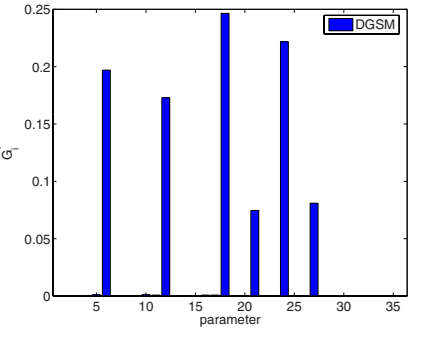


Fig. 6. DGSM, Set of bounds II

4 Conclusions

Accurate parameter estimation of complex dynamical systems can become very challenging as models can contain hundreds or even thousands of parameters. The results of a sensitivity analysis can be used to establish a ranking of importance for the parameters, allowing to fix unessential parameters and to prioritize the most influential ones and facilitating the identification task.

In this contribution, a novel methodology for global sensitivity analysis reducing $(n+2)$ times the number of function evaluations with respect to the well known Sobol' indices was presented. The results obtained by both methods are equivalent and more robust than those provided by the local sensitivities that are strongly dependent on the selection of the nominal point.

Moreover, we have illustrated the importance of properly choosing the bounds for the parameters when using global sensitivities. This aspect is frequently disregarded and can lead to spurious results. Especial attention should be payed to the number of sample points considered for the integration since low convergence may also be the source of inaccurate results.

Acknowledgements

Authors thank the EU ERASysBio and the Spanish Ministry of Science and Innovation (SYSMO project "KOSMOBAC", MEC GEN2006-27747-E/SYS) for financial support.

References

1. Campolongo, F., Cariboni, J., Schoutens, W.: The importance of jumps in pricing european options. Reliability Engineering and System Safety 91(10-11), 1148–1154 (2006)
2. Kucherenko, S., Rodriguez-Fernandez, M., Pantelides, C., Shah, N.: Monte carlo evaluation of derivative based global sensitivity measures. Reliability Engineering and System Safety (in press, 2008), http://dx.doi.org/10.1016/j.ress.2008.05.006
3. Maly, T., Petzold, L.R.: Numerical methods and software for sensitivity analysis of differential-algebraic systems. Applied Numerical Mathematics 20, 57–79 (1996)
4. Rodriguez-Fernandez, M., Mendes, P., Banga, J.R.: A hybrid approach for efficient and robust parameter estimation in biochemical pathways. BioSystems 83, 248–265 (2006)
5. Saltelli, A., Chan, K., Scott, E.M.: Sensitivity Analysis. John Wiley & Sons, Inc., New York (2000)
6. Sobol, I.M.: On quasi monte carlo integrations. Mathematics and Computers in Simulation 47, 103–112 (1998)
7. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation 55, 271–280 (2001)

Improving a Leaves Automatic Recognition Process Using PCA

Jordi Solé-Casals[1], Carlos M. Travieso[2], Jesús B. Alonso[2], and Miguel A. Ferrer[2]

[1] Signal Processing Group
University of Vic
c/ de la Laura, 13, E-08500, Vic, Barcelona, Spain
[2] Department of Signals and Communications, CeTIC,
University of Las Palmas de Gran Canaria
Campus de Universitário Tafira s/n, E-35017, Las Palmas de Gran Canaria, Spain
jordi.sole@uvic.cat,{ctravieso, jalonso, mferrer}@dsc.ulpgc.es

Abstract. In this work we present a simulation of a recognition process with perimeter characterization of a simple plant leaves as a unique discriminating parameter. Data coding allowing for independence of leaves size and orientation may penalize performance recognition for some varieties. Border description sequences are then used, and Principal Component Analysis (PCA) is applied in order to study which is the best number of components for the classification task, implemented by means of a Support Vector Machine (SVM) System. Obtained results are satisfactory, and compared with [4] our system improves the recognition success, diminishing the variance at the same time.

Keywords: Principal Component Analysis, Pattern Recognition, Leaves Recognition, Parameterization, Characteristics selection.

1 Introduction

Recognition of tree varieties using samples of leaves, in spite of its biological accuracy limitations, is s simple and effective method of taxonomy [1]. Laurisilva Canariensis is a relatively isolated tree species, in the Canary Islands, biologically well studied and characterized. Twenty-two varieties are present in the archipelago and have simple and composed regular leaves. Our study takes into account sixteen of the twenty simple leaf varieties, with totals of seventy-five individuals per each one. They have been picked over different islands, pressed (for conservation purposes) and scanned in gray tonalities.

From a biological perspective, attention has to be brought to the fact that emphasis on structural characteristics, which are consistent among individuals of a species, instead of quality parameterization (as color, size or tonality), improves recognition performance. Quality parameterization lack of accuracy is due to the fact of leaves individual variability on the same variety as well on leaf variability on a single plant. Plant age, light, humidity, context behavior or distribution of soil characteristics, among other things, contributes for such anomaly.

In spite of the fact that we may consider several biological parameters, as we have done previously [2], in order to generalize such study, in this paper we have just

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 243–251, 2009.
springerlink.com

considered a border parameterization. This system was classified by Hidden Markov Model (HMM) [3] achieving a success of 78.33% [4].

In this present work, we have improved that previous study using the transformation and reduction of border parameterization using Principal Component Analysis (PCA) [5], and classifying its result with Support Vector Machines (SVM) [6][7]. The rest of this paper presents our proposal.

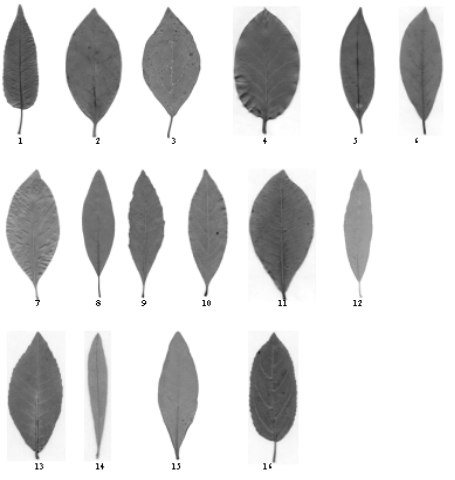


Fig. 1. Images of the 16 varieties of canariensis laurisilva considered for the present study. Images are presented regardless of size.

2 Leaves Database

In order to create a recognition system of different vegetable species it is necessary to build a database. This database should contain the samples of the different species of study. The number of samples will be large enough to, first train the classifier with guarantees and second, test this classifier to assess the results obtained. On top of this, the amount of chosen samples, for each vegetable species must cover the largest amount of shapes and structures that this unique specie can take. In this way, a robust study of the different vegetable species is ensured.

Attending to this reasoning, the sample collection was made at different times of the year, trying in this way to cover all the colors and shapes that the leaves take throughout the four seasons. Besides, a special attention was made to reject those samples that were degraded so that the selected samples were in good condition.

Therefore, this database is composed of 16 classes (see Fig. 1), with 75 samples each one. The images that form the database has been stored in a grey scale using a "jpeg" format (Joint Photographic Experts Group) with Huffman compression. The images have been digitalized to 300 dpi, with 8 bit accuracy.

3 Parameterization System

We have considered just the leaf perimeter. This image is considered without its petiole that has been extracted automatically from the shadow image. Leaves are scanned fixed on white paper sheets, placed more or less on the center, upward (petiole down) and reverse side to scan.

Border determination as (x,y) positioning perimeter pixels of black intensity, has been achieved by processes of shadowing (black shape over white background), filtering of isolated points, and perimeter point to point continuous follow.

3.1 Perimeter Interpolation

As shown in table 1, perimeter size variability induces us to consider a convenient perimeter point interpolation, in order to standardize perimeter vector description. For an interpolating process, in order to achieve reconstruction of the original shape, we may use any of the well known algorithms as mentioned in [8], [9], [10], but a simple control point's choice criterion in 1-D analysis allows for an appropriate performance ratio on uniform control point's number and approximation error for all individuals of all varieties studied.

Table 1. A comparative table of mean error, obtained from a uniform criterion of control point selection and the monotonic way

Class	Mean size	Mean Error	
		Uniform	Monotonic
01	2665.6	9.1474	2.0226
02	1885.1067	3.5651	0.43655
03	2657.68	11.0432	5.3732
04	2845.8133	31.6506	2.8447
05	1994.68	1.8569	0.42231
06	2483.04	0.4425	0.71093
07	2365.2667	9.711	0.68609
08	3265.48	0.4753	0.49015
09	2033.2267	19.7583	3.4516
10	2258.2533	3.9345	2.4034
11	1158.9867	5.4739	1.0286
12	1934	1.3393	0.40771
13	1183.4	1.2064	0.39012
14	981.4	0.2752	0.23671
15	3159.08	11.575	8.8491
16	1973.3733	47.4766	6.6833

The general idea, for such choice, is to consider (x,y) positional perimeter points as (x,F(x)) graph points of a 1-D relation F.

Consideration of y coordinate as y = F(x) is done, because of the way, leaves images are presented in our study: leaves have been scanned with maximum size placed over x ordinate.

For a relation G to be considered as a one-dimensional function, there is need to preserver a correct sequencing definition (monotonic behavior).

That is: A graph,

$$G = \{ i = 1..n, (x_i, y_i) / \; y_i = f(x_i) \} \tag{1}$$

It is the description of a function f if ordinate points $x_i, i = 1..n$ must be such that: $x_i < x_{i+1}, i = 1..n-1$.

We consider then the border relation F as a union of piece like curves (graphs) preserving the monotonic behavior criterion, i.e.

$$F = \bigcup_{j \in J} G_j \tag{2}$$

where: $G_j \subseteq F, \forall \; j \in J$ and $G_j = \{ \alpha_j \in J_j, (x_{\alpha_j}, y_{\alpha_j}) / \; y_{\alpha_j} = f_j \}$,

For convenient sets of index J, J_j and restriction functions $f_j = f_{|\{x_{\alpha_j} | \alpha_i \in J_j\}}$, such that the next point following the last of G_j is the first one of G_{j+1}. Gj graphs are correct f_j functions descriptions.

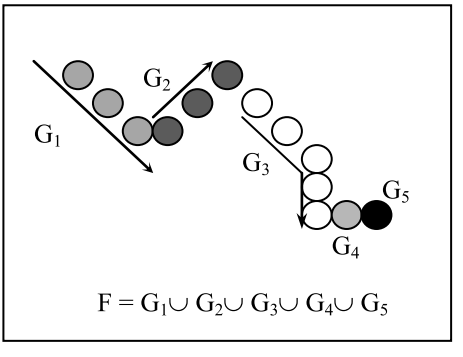


Fig. 2. Example of an F relation decomposed in graphs with a correct function description

Building the Gj sets is a very straightforward operation:

- Beginning with a first point we include the next one of F.
- As soon as this point doesn't preserve monotonic behavior we begin with a new $G_{j+}1$.
- Processes stop when all F points are assigned.

In order to avoid building G_j reduced to singletons, as show in figure 2 (G_4 and G_5) the original F relation may be simplified to preserve only the first point of constant x ordinate series.

Afterwards, spreading of a constant number of points is done proportional to the length of the G_j and always setting in it is first one.

The point's choice criterion mentioned before allows, in two-dimensional interpolation, for taking account on points where reverse direction changes take place. Irregularity, of the surface curve, is taken into account with a sufficient number of interpolating points, as done in the uniform spreading way. Results on table 1 allows for comparison between choice of control points with the criterion motioned before and the uniform one. Such results show the benefit of choosing control points with the monotonic criterion instead of the uniform one.

The 1-D interpolation has been perform using 359 control points, with spline, lineal or closest interpolated point neighborhood, depending on the number of control points present in the decomposed curve. As a reference at 300 dpi a crayon free hand trace is about 5 to 6 points wide.

Table 1 also shows size variability of the different varieties ranging in mean, between 981 pixels for class 14 to 3255 for class 8. With 359 points chosen with the monotonic criterion, all perimeter point vectors have a standard size and errors representation is negligible.

Due to perimeter size variability inside a class, for example in class 15 ranging between 2115 points to 4276 with a standard deviation of about 521, coding of (x,y) control perimeter points have been transformed taking account for size independence.

Considering the following definitions:

Γ the set of n, a fixed number, of control points, $\Gamma = \{X_{i=1..n} / X_i = (x_i, y_i)\}$ Where (x_i,y_i) are point coordinates of control perimeter points.

C_0 the central point of the Γ set: $C_0 = (1/n)(\sum_{i=1..n} x_i, \sum_{i=1..n} y_i)$, $/(x_i, y_i)_{i=1..n} \in \Gamma$, $\beta_i = angle(C_0 X_i X_{i+1})$, $\alpha_i = angle(X_i C_0 X_{i+1})$ angles defined for each interpolating points of Γ.

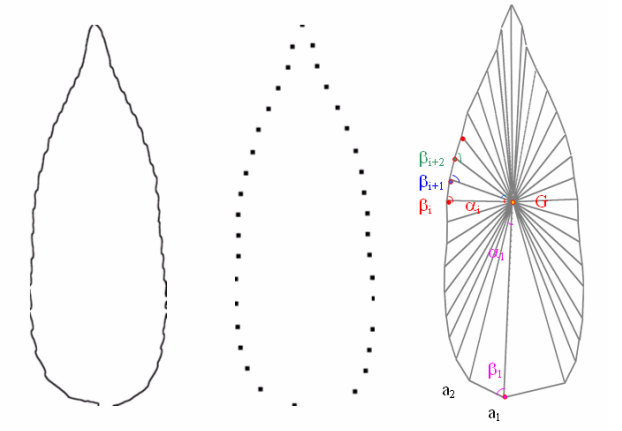


Fig. 3. Example of an angular coding for a 30 control points selection

An example is shown in figure 3. Sequences of (x_i,y_i) positional points are then transformed in sequence of (φ_i, β_i) angular points.

The choice of a starting and a central point accounts for scale and leaf orientation. Placement of both points sets the scale: its distance separation. Relative point positioning sets the orientation of the interpolating shape. Given a sequence of such angles α_i and β_i, it's then possible to reconstruct the interpolating shape of a leaf. Geometrical properties of triangle similarities make such sequence size and orientation free.

4 Reduction Parameters

Principal Components Analysis (PCA) is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [5]. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, i.e. by reducing the number of dimensions, without much loss of information.

PCA is an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms.

In PCA, the basis vectors are obtained by solving the algebraic eigenvalue problem $\mathbf{R}^{\mathbf{T}}\left(\mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{R} = \Lambda$ where $\mathbf{X}$ is a data matrix whose columns are centered samples, $\mathbf{R}$ is a matrix of eigenvectors, and Λ is the corresponding diagonal matrix of eigenvalues. The projection of data $\mathbf{C}_n = \mathbf{R}_n^{\mathbf{T}}\mathbf{X}$ **,** from the original *p* dimensional space to a subspace spanned by *n* principal eigenvectors is optimal in the mean squared error sense.

Another possibility, not presented here, is to use Independent Component Analysis (ICA) [11] [12] instead of PCA. In this case, we obtain independent coordinates and not only orthogonal as in previous case. ICA has been used for dimensional reduction and classification improvement with success [13].

In our problem we have 16 different classes of leaves, and for each class we have 75 different samples, where each one is a two column matrix of 359 points. First column corresponds to interior angles α_i and second column to exterior angles β_i, as explained before (see Fig. 3).

The procedure for applying PCA can be summarized as follows:

1. Subtract the mean from each of the data dimensions. This produces a data set whose mean is zero (**X**).
2. Calculate de covariance matrix ($\mathbf{Cov}_x$)
3. Calculate the eigenvectors e_i and eigenvalues λ_i of $\mathbf{Cov}_x$
4. Order the eigenvectors e_i by eigenvalue λ_i, highest to lowest. This gives us the components in order of significance.

5. Form a feature vector by taking the eigenvectors that we want to keep from the list of eigenvectors, and forming a matrix (**R**) with these eigenvectors in the columns.
6. Project the data to a subspace spanned by these *n* principal components.

5 Classification

For the classification system based on the SVM [6], [7], in order to establishing efficiency, we have calculated error, success and rejected rates on recognition.

Particularly, we have used an implementation of Vapnik`s Support Vector Machine known as SVM light [6], [7] which is a fast optimization algorithm for pattern recognition, regression problem, and learning retrieval functions from unobtrusive feedback to propose a ranking function. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently.

In the next figure, we can see the detection of support vectors and the creation of a boundary, one per each class, because it is a bi-class classifier (see figure 4). In our implementation, we have built a multi-classes classification module, from this SVM light.

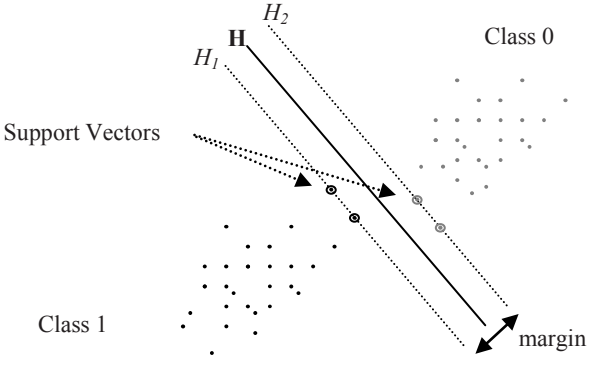


Fig. 4. Separate lineal Hyperplane in SVM

6 Experiments and Results

We did several experiments in order to find the best dimension reduction for leaves automatic recognition. In our experiments we observed that the first column of each sample, that corresponds to the interior angles α_i are not useful for the classification purpose. Hence, we use only exterior angles β_i.

To apply PCA to these values we construct a global matrix with 37 of 75 different samples that we have for each class, arranged in rows, resulting in a 592x359 global matrix. Procedure detailed in Section 4 is applied to this matrix in order to obtain the projected data by using the subspace spanned by 1 to 15 principal components (only odd numbers in our experiments).

Table 2. Results with SVM classifier

Number of Components	Success Rates	Type of kernel	Γ
1	41.53% ± 3.72	Lineal	---
3	44.58% ± 0.33		
5	49.94% ± 7.49		
7	58.62% ± 0.90		
9	62.33% ± 3.40		
11	63.70% ± 0.11		
13	63.43% ± 0.12		
15	69.25% ± 0.33		
1	58.27% ± 0.98	RBF	8×10^{-2}
3	62.23% ± 0.70		0.9
5	78.91% ± 0.33		7×10^{-2}
7	83.69% ± 0.80		0.7
9	83.69% ± 2.01		0.6
11	84.58% ± 0.62		1×10^{-1}
13	86.25% ± 0.45		6×10^{-1}
15	87.14% ± 0.86		7×10^{-2}

Data processed with PCA is used then with the SVM classificatory and results are shown in Table 2. As the methodology of our experiments was a cross-validation method repeating each experiments 10 times, Table 2 shows the obtained average ± typical deviation success rate for different number of PCA components considered in the experiments.

We compare our results with the results obtained in [4], where a HMM of 40 stages was used in the best case, giving a success rate of 78.33% ± 6.06. As can be seen in Table 2, RBF kernel for a SVM system give much better results than lineal kernel whatever the number of components is used. In all of the cases of SVM with RBF kernel and 5 components or more, we outperforms the HMM results in success rate and we diminish the variance as well. The best case is obtained with RBF kernel and a PCA of 15 components, that significantly improves the HMM results.

7 Conclusions

In this present work, we have presented an improvement of an automatic leaves recognition system using Principal Component Analysis and classifying with Support Vector Machines. The transformation and reduction of data contribute to increase its discrimination, from 78.33% using contour parameterization + HMM to 87.14% using contour parameterization + PCA + SVM. The advantage of using PCA is twofold: first, we increase the classification results, and second we diminish the features dimension, giving as a result a less complex classifier. Future work will be done using ICA as an alternative method to PCA for improving results, and other kind of classifiers will be explored.

Acknowledgments

The first author acknowledges support from the Ministerio de Educación y Ciencia of Spain under the grant TEC2007-61535/TCM, and from the Universitat de Vic under the grant R0912.

References

1. Lu, F., Milios, E.E.: Optimal Spline Fitting to Planar Shape. Elsevier Signal Processing 37, 129–140 (1994)
2. Loncaric, S.: A Survey of Shape Analysis Techniques. Pattern Recognition 31(8), 983–1001 (1998)
3. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs (1993)
4. Briceño, J.C., Travieso, C.M., Ferrer, M.A.: Automatic Recognition of Simple Laurisilva Canariensis Leaves, by Perimeter Characterization. In: IASTED International Conference on Signal Processing, Pattern Recognition and its Applications, pp. 249–254 (2002)
5. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer Series in Statistics. Springer, Heidelberg (2002)
6. Cui, G., Feng, G., Shan, S.: Face Recognition Based on Support Vector Method. In: Proceedings of 5th Asian Conference on Computer Vision, pp. 23–28 (2002)
7. Guo, G., Li, S.Z., Kapluk, C.: Face recognition by support vector machines. Image and Vision Computing 19(9-10), 631–638 (2001)
8. Lu, F., Milios, E.E.: Optimal Spline Fitting to Planar Shape. Elsevier Signal Processing 37, 129–140 (1994)
9. Loncaric, S.: A Survey of Shape Analysis Techniques. Pattern Recognition 31(8), 983–1001 (1998)
10. Huang, Z., Cohen, F.: Affine-Invariant B-Spline Moments for Curve Matching. IEEE Transactions on Image Processing 5(10), 824–836 (1996)
11. Jutten, C., Herault, J.: Blind separation of sources, Part 1: an adaptive algorithm based on neuromimetic architecture. Signal Processing 24(1) (July 1991); ISSN:0165-1684
12. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, New York (2001)
13. Poblador, V.S., Moreno, E.M., Solé-Casals, J.: ICA as a preprocessing technique for Classification. In: Puntonet, C.G., Gonzalez Prieto, A. (eds.) ICA 2004. LNCS, vol. 3195, pp. 1165–1172. Springer, Heidelberg (2004)

Author Index