Christian Böhm
Sami Khuri
Lenka Lhotská
M. Elena Renda (Eds.)

LNCS 7451

Information Technology in Bio- and Medical Informatics

Third International Conference, ITBAM 2012
Vienna, Austria, September 2012
Proceedings

Springer

Lecture Notes in Computer Science 7451

Christian Böhm Sami Khuri
Lenka Lhotská M. Elena Renda (Eds.)

Information Technology in Bio- and Medical Informatics

Third International Conference, ITBAM 2012
Vienna, Austria, September 4-5, 2012
Proceedings

Springer

Volume Editors

Christian Böhm
Ludwig-Maximilians-University, Department of Computer Science
Oettingenstraße 67, 80538 München, Germany
E-mail: boehm@dbs.ifi.lmu.de

Sami Khuri
San José State University, Department of Computer Science
One Washington Square, San José, CA 95192-0249, USA
E-mail: sami.khuri@sjsu.edu

Lenka Lhotská
Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics
Technicka 2, 166 27 Prague 6, Czech Republic
E-mail: lhotska@fel.cvut.cz

M. Elena Renda
Istituto di Informatica e Telematica del CNR
Via G. Moruzzi 1, 56124 Pisa, Italy
E-mail: elena.renda@iit.cnr.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-32394-2 e-ISBN 978-3-642-32395-9
DOI 10.1007/978-3-642-32395-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012943467

CR Subject Classification (1998): H.3, H.2.8, H.4-5, J.3, J.1

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Biomedical engineering and medical informatics represent challenging and rapidly growing areas. Applications of information technology in these areas are of paramount importance. Building on the success of ITBAM 2010 and ITBAM 2011, the aim of the third ITBAM conference was to continue bringing together scientists, researchers, and practitioners from different disciplines, namely, from mathematics, computer science, bioinformatics, biomedical engineering, medicine, biology, and different fields of life sciences, so that they can present and discuss their research results in bioinformatics and medical informatics. We hope that ITBAM will always serve as a platform for fruitful discussions between all attendees, where participants can exchange their recent results, identify future directions and challenges, initiate possible collaborative research, and develop common languages for solving problems in the realm of biomedical engineering, bioinformatics, and medical informatics.

The importance of computer-aided diagnosis and therapy continues to draw attention worldwide and has laid the foundations for modern medicine with excellent potential for promising applications in a variety of fields, such as, telemedicine, Web-based healthcare, analysis of genetic information and personalized medicine. Following a thorough peer-review process, we selected 12 long papers and three short papers for the third annual ITBAM conference. The Organizing Committee would like to thank the reviewers for their excellent job. The articles can be found in the proceedings and are divided into the following sections: Medical Data Mining and Information Retrieval; Metadata Models, Prediction, and Mobile Applications; Systems Biology and Data Mining in Bioinformatics. The papers show how broad the spectrum of topics in applications of information technology to biomedical engineering and medical informatics is.

The editors would like to thank all the participants for their high-quality contributions and Springer for publishing the proceedings of this conference. Once again, our special thanks go to Gabriela Wagner for her hard work on various aspects of this event.

June 2012

Christian Böhm
Sami Khuri
Lenka Lhotska
M. Elena Renda

Organization

General Chair

Christian Böhm	University of Munich, Germany

Conference Program Chairs

Sami Khuri	San José State University, USA
Lenka Lhotska	Czech Technical University Prague, Czech Republic
M. Elena Renda	IIT - CNR, Pisa, Italy

Program Committee

Werner Aigner	FAW
Fuat Akal	Functional Genomics Center Zurich, Switzerland
Tatsuya Akutsu	Kyoto University, Japan
Andreas Albrecht	Queen's University Belfast, UK
Lijo Anto Manjaly Antony	Centre for Bioinformatics, University of Kerala, India
Rubén Armañanzas Arnedillo	University of the Basque Country, Spain
Peter Baumann	Jacobs University Bremen, Germany
Balaram Bhattacharyya	Visva-Bharati University, India
Christian Blaschke	Bioalma Madrid, Spain
Andreas M. Boehm	Rudolf Virchow Center for Experimental Biomedicine, Germany
Veselka Boeva	Technical University of Plovdiv, Bulgaria
Gianluca Bontempi	Université Libre de Bruxelles, Belgium
Roberta Bosotti	Nerviano Medical Science s.r.l., Italy
Rita Casadio	University of Bologna, Italy
Sònia Casillas	Universitat Autònoma de Barcelona, Spain
Kun-Mao Chao	National Taiwan University
Vaclav Chudacek	Czech Technical University in Prague, Czech Republic
Coral del Val Muñoz	University of Granada, Spain
Hans-Dieter Ehrich	Technical University of Braunschweig, Germany
Mourad Elloumi	University of Tunis, Tunisia
Maria Federico	University of Modena and Reggio Emilia, Italy
Pedro Fernandes	Inst.Gulbenkian de Ciência, Portugal

Christoph M. Friedrich	University of Applied Sciences Dortmund, Germany
Xiangchao Gan	University of Oxford, UK
Alejandro Giorgetti	University of Verona, Italy
Alireza Hadj Khodabakhshi	Simon Fraser University, Canada
Volker Heun	Ludwig-Maximilians-Universität München, Germany
Chun-Hsi Huang	University of Connecticut, USA
Lars Kaderali	University of Technology Dresden, Germany
Alastair Kerr	University of Edinburgh, UK
Michal Krátký	Technical University of Ostrava, Czech Republic
Vaclav Kremen	Czech Technical University in Prague, Czech Republic
Josef Küng	University of Linz, Austria
Gorka Lasso	CICbioGUNE, Derio, Spain
Roger Marshall	Plymouth State University, USA
Elio Masciari	ICAR-CNR, Università della Calabria, Italy
Henning Mersch	RWTH Aachen University, Germany
Aleksandar Milosavljevic	Baylor College of Medicine, USA
Jean-Christophe Nebel	Kingston University, London, UK
Vit Novacek	National University of Ireland, Galway, Ireland
Nadia Pisanti	University of Pisa, Italy
Cinzia Pizzi	Università degli Studi di Padova, Italy
Clara Pizzuti	Institute for High Performance Computing and Networking (ICAR)-National Research Council(CNR), Italy
Meikel Poess	Oracle Corporation, Redwood Shores, CA, USA
Nicole Radde	Universität Stuttgart, Germany
Stefano Rovetta	University of Genova, Italy
Cristina Rubio-Escudero	University of Seville, Spain
Nick Sahinidis	Carnegie Mellon University, USA
Roberto Santana	University of the Basque Country (UPV/EHU), Spain
Kristan Schneider	University of Vienna, Austria
Kathleen Steinhofel	King's College London, UK
A Min Tjoa	Vienna University of Technology, Austria
Paul van der Vet	University of Twente, The Netherlands
Roland R. Wagner	University of Linz, Austria
Viacheslav Wolfengagen	Institute JurInfoR-MSU, Russia
Borys Wrobel	Polish Academy of Sciences, Poland
Filip Zavoral	Charles University in Prague, Czech Republic
Songmao Zhang	Chinese Academy of Sciences, China
Qiang Zhu	The University of Michigan, USA
Frank Gerrit Zoellner	University of Heidelberg, Germany

Table of Contents

Session 1

Intelligent Data Acquisition and Scoring System for Intensive Medicine 1
Filipe Portela, Manuel Filipe Santos, José Machado, Álvaro Silva, Fernando Rua, and António Abelha

Data Mining in the Study of the Chronic Obstructive Pulmonary Disease 16
Maribel Yasmina Santos, Jorge Cruz, and Artur Teles de Araújo

Practical Problems and Solutions in Hospital Information System Data Mining 31
Miroslav Bursa, Lenka Lhotska, Vaclav Chudacek, Jiri Spilka, Petr Janku, and Martin Huser

Session 2

Using Generic Meta-Data-Models for Clustering Medical Data 40
Dominic Girardi, Michael Giretzlehner, and Josef Küng

A Mobile Based Authorization Mechanism for Patient Managed Role Based Access Control 54
Cátia Santos-Pereira, Alexandre B. Augusto, Manuel E. Correia, Ana Ferreira, and Ricardo Cruz-Correia

Care@HOME: A Mobile Monitoring System for Patient Treatment and Blood Pressure Tracking 69
Mersini Paschou, Efrosini Sourla, George Basagiannis, Evangelos Sakkopoulos, and Athanasios Tsakalidis

Session 3

An Integrative Clustering Approach Combining Particle Swarm Optimization and Formal Concept Analysis 84
Anna Hristoskova, Veselka Boeva, and Elena Tsiporkova

Link Prediction Approaches for Disease Networks 99
Francesco Folino and Clara Pizzuti

Toward a Semantic Framework for the Querying, Mining and Visualization of Cancer Microenvironment Data 109
Michelangelo Ceci, Fabio Fumarola, Pietro Hiram Guzzi, Federica Mandreoli, Riccardo Martoglia, Elio Masciari, Massimo Mecella, and Wilma Penzo

Session 4

Argumentation to Represent and Reason over Biological Systems 124
Adam Wyner, Luke Riley, Robert Hoehndorf, and Samuel Croset

The Use of Design Specificity in Standardized Mean Difference for Analysis of High throughput RNA Interference Screens 139
Karol Kozak

Toward a Translational Medicine Approach for Hypertrophic Cardiomyopathy 151
Catia M. Machado, Francisco M. Couto, Alexandra R. Fernandes, Susana Santos, and Ana T. Freitas

Poster Session

Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation 166
Andreas Holzinger, Reinhold Scherer, Martin Seeber, Johanna Wagner, and Gernot Müller-Putz

The Database of the Cardiovascular System Related Signals 169
Jan Havlík, Lucie Kučerová, Imrich Kohút, Jan Dvořák, and Vratislav Fabián

Patient Monitoring Using Bioimpedance Signal 171
Jan Havlík, Ondřej Fousek, and Miroslav Ložek

Author Index 173

Intelligent Data Acquisition and Scoring System for Intensive Medicine

Filipe Portela[1], Manuel Filipe Santos[1], José Machado[2],
Álvaro Silva[3], Fernando Rua[3], and António Abelha[2]

[1] Centro Algoritmi, University of Minho, Guimarães, Portugal
{cfp,mfs}@dsi.uminho.pt
[2] CCTC, University of Minho, Braga, Portugal
{jmac,abelha}@di.uminho.pt
[3] Serviço de Cuidados Intensivos, Centro Hospitalar do Porto, Hospital Santo António, Portugal
moreirasilva@clix.pt
fernandorua.sci@hgsa.min-saude.pt

Abstract. In a critical area as is Intensive Medicine, the existence of systems to support the clinical decision is mandatory. These systems should ensure a set of data to evaluate medical scores like is SAPS, SOFA and GLASGOW. The value of these scores gives the doctors the ability to understand the real condition of the patient and provides a mean to improve their decisions in order to choose the best therapy for the patient. Unfortunately, almost all of the required data to obtain these scores are recorded on paper and rarely are stored electronically. Doctors recognize this as an important limitation in the Intensive Care Units. This paper presents an intelligent system to obtain the data, calculate the scores and disseminate the results in an online, automatic, continuous and pervasive way. The major features of the system are detailed and discussed. A preliminary assessment of the system is also provided.

Keywords: Intensive Medicine Scores, Real Time, Pervasive System, Scoring System, SAPS, SOFA, GLASGOW.

1 Introduction

During the last three decades, several physiological-based prognostic models have emerged [1]. In 1980, the severity scoring systems were introduced in the Intensive Care Units (ICUs) [2]. Intensive-care medicine score systems serve to quantify the severity of diseases and to characterize patient groups on the basis of objective criteria [1]. The scores describe patient severity by adding up points. As main objectives, they: assess the prognosis; establish the amount of treatment required; provide information on the prognosis of patients; indicate the efficacy of therapeutic interventions; and serve for the stratification in clinical studies and workload [2], helping doctors in the decision process. In the ICU a high number of scoring systems are used that can be grouped in: diseases, patient and universally mortality prediction.

This work has been developed in the context of INTCare research project [3], an intelligent decision support system that makes use of online learning to predict the

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 1–15, 2012.

organ failure and the outcome of the ICU patients. The work developed was only focused on the scores that have variables that are used by Decision Support System, i.e., SAPS, SOFA and Glasgow. In the context of INTCare, a new set of data had to be acquired in order to induce Data Mining (DM) models able to predict the organ failure and the outcome of the patients. These scores rarely were stored in an electronic way. As a normal procedure, the nurses / doctors calculate these scores by consulting the patient's bed side monitors, interpreting the lab results and reading admission documents in an offline mode. The existence of multiple data sources for each organic system [4] difficult the evaluation of the variables by the human and, consequently, can delay / interfere negatively in the decision making process.

For the development of INTCare system this situation was unaffordable. The inexistence of all data in an electronic mode and in real time made impossible the construction of prediction models. Those limitations dictated the development of a completely new scoring system to collect the data and to score the measures in real-time. A lot of research has been done to understand how it would be possible to acquire and prepare the necessary data. An intelligent agent based approach has been followed to perform crucial tasks such as the automatic data acquisition and the online scoring. The main goal of the project was the concentration of the decision support tasks in a single platform.

According to Keegan [1] there are several limitations inherent to the ICU prognostic models: i) errors in data collection and data entry, flaws in development and validation of the models; ii) variables used to make predictions may not be easily measured; and some laboratory values may not be routinely obtained. The lack of standardization in obtaining values leads to missing data and to the fault of important prognostic variables, compromising the performance of the models [5]. The empirical experience accumulated in this field of acting allowed the development of a system able to overcome the limitations described above. This system has been deployed in the ICU of the Centro Hospitalar do Porto (CHP), Porto, Portugal.

This paper presents an intelligent scoring system to feed DM models as a way to predict the organ failure and the outcome of the ICU patients. Chapter one introduces the problematic and the subject, the second chapter makes an overview of the situation and explains the main concepts. Chapter three present the data manipulation process. Chapters four and five will present the scoring system and the results achieved. The preliminary results of a study on the technological acceptance are presented. Finally, some conclusions will be done about the work and the future work.

2 Background

It is common sense that in the ICU a set of clinical measures should be used regularly. Many of the scoring processes are executed manually and require a high level of human intervention. Currently there are systems able to collect all data produced during the day and execute some scoring. An example of this is the system presented by Shabot [6] where "Three measures of severity of illness are automatically calculated for each adult ICU patient on admission and again daily". When the scores are calculated manually the value considered are not the worst among the measured values but the worst value registered by someone. This configures a major limitation.

For a more accurate decision it is desirable to develop a system able to calculate the scores automatically according to the data captured from the patients along the day in real time.

2.1 Intensive Care Scores

Scores are integrated in the diagnosis-related groups [7] and can be used, for example, to predict the outcome [8]. In the ICU of CHP, the most used scores are: SOFA, SAPS II and Glasgow.

Sepsis-related Organ Failure Assessment (SOFA) is used to daily score, as objectively as possible, the degree of organ dysfunction/failure of a patient [9]. SOFA considers the worst value occurred along the day for the calculation of a score of 0 (no failure) to 4 (severe failure) for each organ: respiratory, renal, cardiovascular, neurologic, coagulation and hepatic [4].

Simplified Acute Physiology Score II (SAPS II) is an evolution of SAPS and provides an estimation of the risk of death without having to specify a primary diagnosis. SAPS II scores are converted into a probability of hospital mortality [10] making use of logistic regression analysis [4]. It uses twelve physiologic variables more age, admission type and the presence of metastatic or haematological cancer or AIDS [4].

More recently SAPS III has been developed. SAPS III includes a set of new variables, prefacing a total of 20 variables [5]: socio-demographics, chronic conditions, diagnostic information, physiological derangement at ICU admission, number and severity of organ dysfunctions, length of ICU and hospital stay, and vital status at ICU and hospital discharge. SAPS III is a model that was developed to assess severity of illness and to predict vital status at hospital discharge based on ICU admission data [11].

Glasgow Coma Score (GCS) [12] describes the patient's level of consciousness. GCS is scored between 3 and 15, where 3 is the worst value, and 15 the best. This score can’t be automatically calculated because it requires human observation. It can be calculated several times along the day and is composed by three parameters: best eye response, best verbal response and best motor response.

2.2 Intcare System

INTCare [3] is a research project whose main goal is to develop an Intelligent Decision Support System to, automatically and in real-time, predict the organ failure and patient outcome by means of data mining models [13]. INTCare system is divided in fours subsystem: data acquisition, knowledge management, inference and interface [14]. The DM models are fed from several data sources including ICU scores, in this case the SOFA values and some data from SAPS. The results described in this paper were obtained through the INTCare system. INTCare has been implemented in terms of intelligent agents able to react and / or reasoning that work together to obtain the ICU scores autonomously [15][16].

3 Data Extraction, Tranformation and Loading

Extraction, Transformation and Loading (ETL) is the process responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse [17]. For scoring purposes, the following systems and data were considered:

- Bedside Monitors - vital signs (VS);
- Electronic Health Record (EHR) - patient admission values;
- Electronic Nursing Record (ENR) – Hourly values (Fluid balance, Glasgow);
- Drugs System (DS) – therapeutic plan;
- Laboratory Results (LR) – Blood and blood gas exam results;

A set of agents of the INTCare data acquisition subsystem [14] are in charge of the tasks associated to ETL [15]. Fig. 1 illustrates the overall ETL process. The data acquisition system is composed by a set of agents in charge of collecting data from several data sources automatically (bedside monitors, drugs system, EHR, laboratory and ENR). After this, the data is corrected and stored in order to be used by the scoring system. Finally, the data collected is used to calculate the ICU measures. This process is explained in detail in the next lines making use of the agent's paradigm.

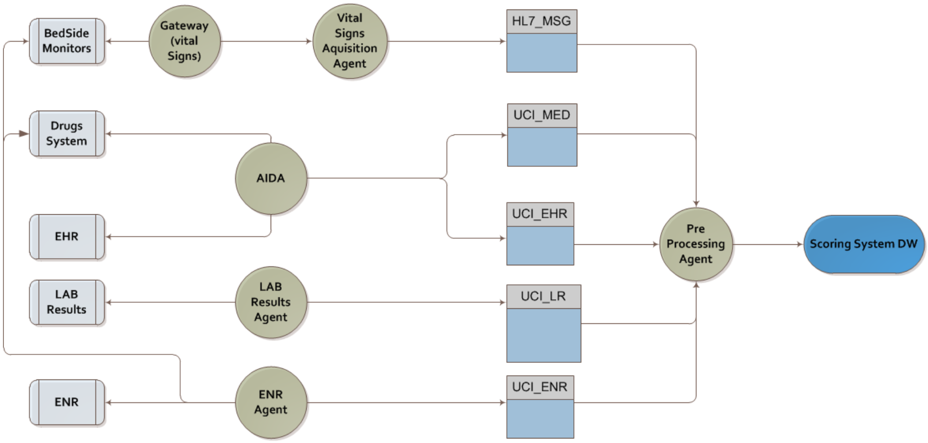


Fig. 1. ETL Process

3.1 Extraction Process

Each score uses a set of different variables however, some of them are common. Table 1 presents the variables collected in real-time. For each variable indicates its correspondent data source and the way it is acquired (automatically or manually).

Table 1. Scores Variables data source and acquisition type

Variable	Score	Data Source	Acquisition
Acute infection	SAPS III	HER	Manual
Admission Glasgow Score	Glasgow	EHR	Automatic
Age	SAPS II	EHR	Automatic
Anatomical surgery site	SAPS III	EHR	Manual
Bicarbonate	SAPS II	LR	Automatic
Bilirubin	SOFA, SAPS II, SAPS III	LR	Automatic
Chronic diseases	SAPS II	EHR	Automatic
Co-Morbidities	SAPS III	EHR	Automatic
Creatinine	SOFA, SAPS II, SAPS III	LR	Automatic
Eye response	Glasgow	ENR	Manual (hour)
Glasgow coma scale	SAPS II, SAPS III	EHR	Automatic
Glasgow coma scale	SOFA	ENR	Manual (hour)
Heart rate	SAPS II, SAPS III	VS | ENR	Automatic
Hydrogen ion	SAPS III	LR	Automatic
Intra-hospital location	SAPS III	EHR	Automatic
Length of stay	SAPS III	EHR	Automatic
Leukocytes	SAPS III	LR	Automatic
Mean Arterial Pressure	SOFA	VS | ENR	Automatic
Mechanical ventilation	SAPS II	ENR	Automatic
Motor response	Glasgow	ENR	Manual (hour)
PaO_2/FiO_2	SOFA, SAPS II, SAPS III	LR	Automatic
Platelets	SOFA, SAPS III	LR	Automatic
Potassium	SAPS II	LR	Automatic
Reason(s) for admission	SAPS III	EHR	Automatic
Serum Urea or BUN	SAPS II	LR	Automatic
Sodium	SAPS II	LR	Automatic
Surgical status	SAPS III	EHR	Automatic
Systolic BP	SAPS II, SAPS III	VS | ENR	Automatic
Temperature	SAPS II, SAPS III	VS | ENR	Automatic
Type of admission	SAPS II, SAPS III	EHR	Automatic
Urine Output	SAPS II, SAPS III	ENR	Manual (hour)
Use of major therapeutic	SAPS III	EHR	Manual
Vasopressors	SOFA	DS | ENR	Automatic
Verbal response	Glasgow	ENR	Manual (hour)
White Blood Cell	SAPS II	LR	Automatic

Next, each agent responsible to collect data from the data sources will be explained in terms of its actions and variables instantiated.

Gateway (a_{gat}) operates in real-time and is responsible for capturing the vital signs data from bedside monitors. These data are packed into Health Level Seven (HL7) [18] messages and sent to the Vital Signs Acquisition Agent. Gateway collects information, in average, once at each eight minutes and restarts at each hour to ensure that no communications failures compromise the system.

a_{gat} { Blood Pressure (systolic, mean), temperature, heart rate}

Vital Signs Acquisition (a_{vsa}) don't collects any data, only is used to, in real-time, extract information blocks, splitting the HL7 message and storing them into the database.

ENR Agent (a_{enr}) is associated to the Electronic Nursing Record. It was developed to allow the introduction of ICU information in an electronic mode, being responsible to capture the clinical data from the doctors and nurses (ICU Staff) [3].

a_{enr} { Urine Output, Temperature, Eye Response, Glasgow Coma Scale, Mechanical Ventilation, Motor Response, Verbal Response, Blood Pressure (Systolic, Mean), Temperature, Heart Rate}

LR (a_{lr}) is responsible for capturing the clinical data from the lab results, i.e. blood and blood gas exams. The ENR agent requests to (a_{lr}), every five minutes, new results from the laboratory. This agent verifies if there are new results from a patient and, if so, it stores them into the database.

a_{lr}{ Bicarbonate, Bilirubin, Creatinine, Hydrogen ion, Leukocytes, PaO2, FiO2, Platelets, Potassium, Serum Urea, BUN, Sodium, Vasopressors, White Blood Cell}

AIDA (a_{ada}) is an agency to archive and to disseminate medical exams and results, earlier implemented at the hospital[19]. In this case, it supplies the patient admission data.

a_{ada} {Acute infection, Admission Glasgow Score (Hospital and Service), Age, Anatomical surgery site, Chronic diseases, Co-Morbidities, Intra-hospital location, Length of stay, Reason(s) for admission, Surgical status, Type of admission, Use of major therapeutic}

Pre-processing (a_{pp}) agent is responsible for the correct linking of all the values in order to create a valid medical record for each patient [3]. It is in charge of solving some data acquisition problems [20] and prepare the data to the scoring system.

3.2 Transforming Process

This process is the most important for the scoring system. In this stage all data collected will be processed, i.e., validated and transformed according to the scoring scales. The process is totally ensured by the pre-processing agent and requires a correct acquisition of the data. The vital signs data, after been collected, need to be validated, due to the errors that normally occur in ICU. All data collected will be automatically validated [21] according to the ranges defined in ICU (Table 2). Those values are also subject of a manual validation after the automatic validation. If some values are out of range, they can be corrected by the nurses. The data validation process is quite simple: every time a value is collected by the agent a trigger is executed (1).

Table 2. ICU Data ranges

Vital Sign	*Min*	*Max*
Blood Pressure (BP)	0	300
Temperature (Temp)	35	45
Respiratory Rate (RR)	0	40
Heart Rate (HR)	0	250

```
If value is null
    delete row;                                    (1)
If value >= min and value <= max
    move row_data to table "real_data"
    set valid_data = true
```

After having all data validated, a set of procedures / functions or triggers are executed to prepare the data to be used in the scores.

The next process is a typical process of transformation and cataloguing of the variables according to the scores table. Table 3 is an example of a scoring table that is stored in the database to assist in the transforming process. This table contains the identification of the score, the variable measured, the possible values, the minimum and maximum values allowed for this variable and the points associated to each possible value. For each value collected a new analysis is done. The value will be verified and catalogued according to its importance / significance. The respective score result will be assigned after querying the database. In the case of numeric variables, the value collected is evaluated according to the min and max defined to each possibility. For example, in the case of bilirubin, three different types of measures are used, one for SAPSII, another one for SAPSIII and a last one for SOFA. For each case, the scale (min and max) varies and, according to the score in study the respective point will be associated to the value collected. If the variable collected is based in text type the scoring process will be based on the expression (definition). For example, in the case of Glasgow, the eyes will be evaluated taking into account the expression / reaction (absent to pain, to speech, spontaneous) stored in the database and the correct point (1, 2, 3, 4) will be assigned, according to the expression collected. The main difference among SOFA, GLASGOW and SAPS scoring formula is the time / date used. In the case of SAPS, only is used the worst value collected along the first 24 hours. In the other cases, the score is obtained at the end of each day and the worst values of the day are used.

The next formulas (2-4) present some examples of the data processing. Each variable has different configurations according to the score id. An example of each case mentioned before (number and text) is presented.

Table 3. Scores table (example)

Score ID	*Variable ID*	*Definition*	*Min*	*Max*	*Point*
GLASGOW	Eye	Absent			1
GLASGOW	Eye	To pain			2
GLASGOW	Eye	To speech			3
GLASGOW	Eye	Spontaneous			4
SAPSII	Admission Type	Schedule Chirurgic			0
SAPSII	Admission Type	No chirurgic			6
SAPSII	Admission Type	Urgency Chirurgic			8
SAPSII	AGE	< 40	0	39	0
SAPSII	AGE	[40 ; 59]	40	59	7
SAPSII	AGE	[60 ; 69]	60	69	12
SAPSII	AGE	[70 ; 74]	70	74	15
SAPSII	AGE	[75 ; 79]	75	79	16
SAPSII	AGE	> 80	80	120	18
SAPSII	Bilirubin	<4	0	3,9	0
SAPSII	Bilirubin	4 a 5,9	4	5,9	4
SAPSII	Bilirubin	>= 6	6	9999	8
SAPSIII	Admission Time	<14	0	13	0
SAPSIII	Admission Time	[14 ; 27]	14	27	6
SAPSIII	Admission Time	>=28	28	9999	7
SAPSIII	Bilirubin	<2	0	1,9	0
SAPSIII	Bilirubin	[2 ; 5,9]	2	5,9	4
SAPSIII	Bilirubin	>= 6	6	9999	8
SOFA	Bilirubin	< 1,2	0	1,1	0
SOFA	Bilirubin	1,2-1,9	1,2	1,9	1
SOFA	Bilirubin	2,0-5,9	2	5,9	2
SOFA	Bilirubin	6,0-11,9	6	11,9	3
SOFA	Bilirubin	>=12	11,9	9999	4
SOFA	Creatinine	< 1,2 or < 110	0	1,1	0
SOFA	Creatinine	1,2-1,9 or 110-170	1,2	1,9	1
SOFA	Creatinine	2,0-3,4 or 171-299	2	3,4	2
SOFA	Creatinine	3,5-4,9 or 300-440	3,5	4,9	3

δ – value collected
Ω – min score value
β – max score value
γ – scoring point
ρ – score id
Θ – data source variable id
Щ – score variable id
§ - variable definition

Δ – date (value collected)
φ – date (today)
μ – admission day + 1
θ – patient id
¥ – data source table
Ͼ – score final table
Ͽ – scores table

In case of number (example):

```
For each δ                                                              (2)
Insert into Ͼ (
select γ, φ, θ, Ꚗ, ρ from ¥, Ͽ
where δ >= Ω or δ <= β and Δ = φ and ρ = 'SOFA' and Ꚗ = Θ);
Select max(γ), θ, φ, Ꚗ from Ͼ where ρ = 'SOFA' and Ꚗ = Θ
group by
θ, φ, Ꚗ
```

In case of text (example):

```
For each δ                                                              (3)
Insert into Ͼ (
select γ, φ, θ, Ꚗ, ρ from ¥, Ͽ
where δ = Ͽ and Δ = φ and ρ = 'GLASGOW' and Ꚗ = Θ);
Select max(γ), θ, φ, Ꚗ from Ͼ where ρ = 'GLASGOW' and Ꚗ = Θ
group by
θ, φ, Ꚗ
```

In case of SAPS (text example): (4)

```
For each δ
Insert into Ͼ (
select γ, φ, θ, Ꚗ, ρ from ¥, Ͽ
where δ = Ͽ and Δ = μ and ρ = 'SAPSII' and Ꚗ = Θ);
Select max(γ), θ, φ, Ꚗ from Ͼ where ρ = 'SAPSII' and Ꚗ = Θ
group by
θ, φ, Ꚗ
```

In some cases it is necessary to do transformations before the scoring point is associated as is the case of the variable pao2 / fio2. In this case, the values are normally obtained separately and only then the calculation measure is applied: Pao2 / (Fio2 / 100). Another example is the SOFA cardiovascular, where two different types of measure are used: Mean Blood Pressure and Vasopressors. For this case, all variable possibilities will be evaluated and, in the final, only the worst score will be considered. SAPS II and SAPS III also have similar situations, in all the cases the same method is applied.

3.3 Loading Process

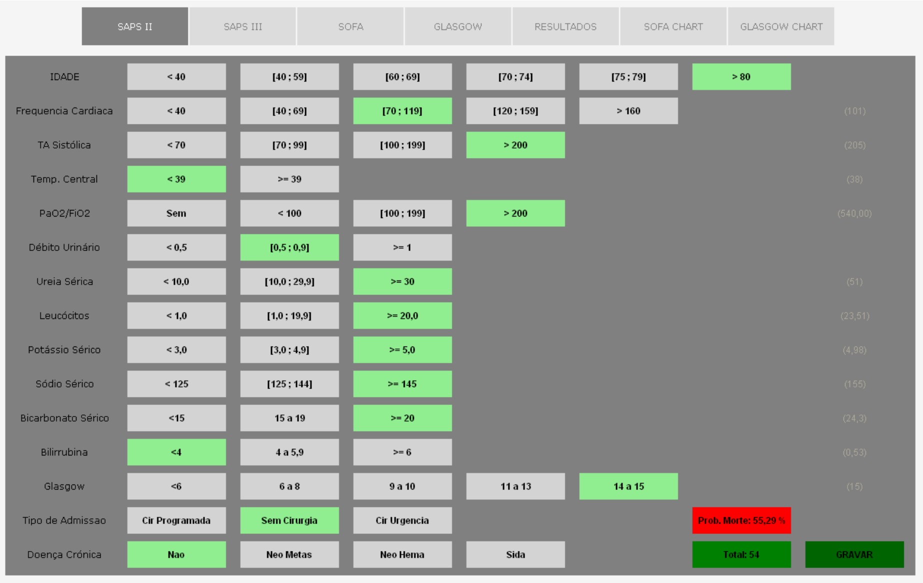
This is the last ETL process. Here, all transformed data are loaded into the data warehouse (DW). Then, all of those data are prepared to be interpreted by the scoring system, allowing for the calculation of the final scores of each measure: SOFA, SAPS II, SAPS III, and GLASGOW.

4 Scoring System

The Scoring System (SS) is integrated in the Electronic Nursing Record (ENR). The ICU staff can consult the results through this application. This application is also used for registering some values that require a human observation like is Glasgow and some SAPS parameters. The next figure (Fig. 2) presents a print screen of a score's view for SAPS II. The other three scores' views are similar. Being this a touch-screen interface, the user can permute among the SCORES sliding the views.

The idea is to give to the ICU staff the better way for consulting the real condition of the patients whenever they want. It is possible to verify the results obtained for each parameter and also the parameters not evaluated until the moment. In the case of the SAPS, the result presented is the worst of the values obtained during the firsts 24 hours. In the other cases (SOFA and Glasgow) the result presented is the worst of the results along the day captured until the actual moment.

The user interface is simple and easy to understand. All the variables of the scale are disposed in a page where, at the left side, are the measure variables names (age, temperature, blood pressure) and at the right the worst absolute value collected for that variable. In the middle of the page are disposed all of the possible results for each score variable in agreement with the score scale. For example, in the first line of the SAPS II score a label for the age is positioned in the left side, in the middle are presented the six possibilities for the age (table 3) and at right the patient age. When a result is obtained the box with the correspondent result will be automatically highlighted in green (eg. 82 years = last box green). With this option the user can quickly understand what are the worst values for the patient for a specific variable.



IDADE	< 40	[40 ; 59]	[60 ; 69]	[70 ; 74]	[75 ; 79]	> 80	
Frequencia Cardiaca	< 40	[40 ; 69]	[70 ; 119]	[120 ; 159]	> 160		(101)
TA Sistólica	< 70	[70 ; 99]	[100 ; 199]	> 200			(205)
Temp. Central	< 39	>= 39					(38)
PaO2/FiO2	Sem	< 100	[100 ; 199]	> 200			(540,00)
Débito Urinário	< 0,5	[0,5 ; 0,9]	>= 1				
Ureia Sérica	< 10,0	[10,0 ; 29,9]	>= 30				(51)
Leucócitos	< 1,0	[1,0 ; 19,9]	>= 20,0				(23,51)
Potássio Sérico	< 3,0	[3,0 ; 4,9]	>= 5,0				(4,98)
Sódio Sérico	< 125	[125 ; 144]	>= 145				(155)
Bicarbonato Sérico	<15	15 a 19	>= 20				(24,3)
Bilirrubina	<4	4 a 5,9	>= 6				(0,53)
Glasgow	<6	6 a 8	9 a 10	11 a 13	14 a 15		(15)
Tipo de Admissao	Cir Programada	Sem Cirurgia	Cir Urgencia			Prob. Morte: 55,29 %	
Doença Crónica	Nao	Neo Metas	Neo Hema	Sida		Total: 54	GRAVAR

Fig. 2. SAPS II Score System

To correct wrong values or add the values in fault is simple. Using the touch screen, the user only needs to click in the correct result and automatically the result is filled or actualized. Although the SS has an automatic saving system the final results require a manual validation. This option ensures that only correct values are stored and they are correctly associated to the patient. The results appear in the system moments after they are collected. Whenever a different result is recorded, the scores are refreshed according to the new data. Taking into accounting the Table 1, it is possible to verify that many of the values are collected automatically (eg. values provided by the laboratory and the bedside monitors). However, in the case of GLASGOW and SAPS III, there is a set of variables that needs to be inserted manually (e.g. urine output, Glasgow). Operation (5) corresponds to the calculation performed to obtain the final results by day and by score.

ρ – score id
Θ – data source variable id
Щ – score variable id
θ – patient id
Δ – date (value collected)
φ – date (today)
Ͼ – score final table

```
For each score and day
    finalScore(ρ, φ) = 0
    For i= 1 to count(Щ)
        Score(i) = (Select max (γ)                                    (5)
        from Ͼ where ρ = 'SCORE' and Щ = 'VariableID(i)' and Δ = φ
        group by
        θ, φ, Щ)
        finalScore(ρ, φ) = finalScore(ρ, φ) + Score(i)
    next;
```

The final score is calculated and presented only when all variables are filled. In the case of the SAPS, the results obtained are also used to predict the death of the patient. SAPS scores are calculated making use of the formulas developed by the field researchers [5, 10] and uses the SAPS II or SAPSS III final results. Each SAPS score uses a different formula to obtain the Predicted Death Rate (PDR). PDR is obtained using the equation 6:

$$PDR = e(Logit)/(1+e(Logit)) \tag{6}$$

SAPS II Logit formula is:

$$Logit = -7{,}7631+0{,}0737*(SAPS\ II)+0{,}9971*\ln((SAPSII)+1) \tag{7}$$

In case of SAPSIII the formula is:

$$Logit = -32.6659 + \ln(SAPS\ III + 20.5958) \times 7.3068 \tag{8}$$

5 Results

A new concept for scores visualization was introduced, based in an hourly and continuous observation of the scores' results. The results are presented both in a grid of values and in a chart format. This is only possible due to a continuous and real-time execution of the data acquisition and data processing. The data processing performed for each hour is the same as that performed for the day (equation 5). The difference is in the value assigned to the φ parameter. If no data is collected for a particular hour, the last result obtained will be considered. This occurs often, principally in the cases where the laboratory results are used (e.g. Creatinine, bilirubin, pao2/fio2). The exams are normally performed in average two/three times a day. When a significant number of values exist for the same hour, as is the case of the vital signs, only the worst value of the hour, correctly collected, will be considered (9). In the hourly approach, the scoring system will be executed using the values obtained until that point.

ρ – score id
Θ – data source variable id
Щ – score variable id
Щn – score variable id (number)
θ – patient id
γ – scoring point

Δ – date (value collected)
φ – date (today)
† – date (hour of today)
ϒ – hour (value collected)
Ͼ – score final table

```
for each score variable
select Щ, max(γ), φ, †, ϒ
from Ͼ
where Δ = φ and θ = 'patient id'
and
ρ = 'SCORE'
and Щ = 'Score Variable'
group by Щ, †, ϒ
order by † asc;
then →
    hNow = 1
    For kj = hNow To ϒ
        If φ is not null and kj = ϒ then
            valueHour(kj, Щn) = max(γ)                    (9)
        ElseIf kj > 0 Then
            valueHour(kj, Щn) = valueHour(kj - 1, Щn)
        End If
        End If
    Next
    hNow = ϒ + 1
next;
```

A different chart is associated to each score. Each chart is composed by the variables that make up the score and the results obtained through the process presented before (9). In the case of SOFA, the chart has seven variables (Neurologic, Respiratory, Coagulation, Liver, Hepatic, Cardiovascular and Total). In Fig. 3 is possible observe the interface of the system developed. For a better comprehension of the values, the scales considered are the same and range between 0 and 24. Only the Sofa total can have values within this range, the other variables (neuro, coag, resp, hepatic, renal, cardio) range from 0 to 4.

These charts are included in the ENR platform and can be consulted in two different ways: hourly (along the 24 hours) and by day (since the admission day). The user can also consult all variables in simultaneous (default) or can isolate a particular variable for a fine-grained study. In the same platform, the ICU staff can record, validate and consult the scores (fig 2).

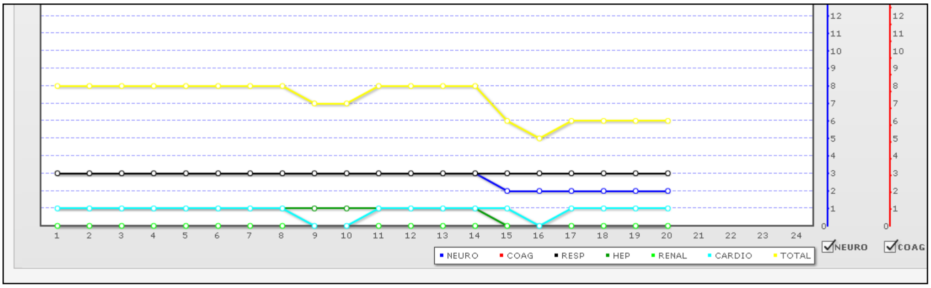


Fig. 3. SOFA score chart

Since its inception in last March, the scoring system was used in more than 20 patients. In order to understand the level of technological acceptance of the system, a questionnaire with ten questions has been conceived. From the ten questions, five are dedicated to the scoring system. 15% of the ICU professionals have been asked to respond to the questionnaire. The evaluation scale was defined around 5 different levels of agreement:

1) In complete disagreement;
2) Disagrees;
3) Agrees;
4) Satisfactory agreement;
5) Complete agreement.

The questions were grouped into two different groups: the functional aspects and the technical aspects. The five questions concerning the scoring system are:

1 An efficient consulting of information for nursing decision support is allowed?
2 An efficient consulting of information for medical decision support is allowed?
3 A proactive performance of the professionals is enhanced by the system?
4 Is the access to the information, in terms of speed and availability, adequate to the needs?
5 Is the access to the system easy and secure?

The Table 4 presents the results obtained, in terms of the percentage of the answers for each question and each evaluation level.

Table 4. Questionnaire result (%)

Query	1	2	3	4	5	Query	1	2	3	4	5
Functional characteristics						**Technical characteristics**					
1	0,00	0,00	0,00	16,67	83,33	3	0,00	0,00	0,00	66,67	33,33
2	0,00	0,00	0,00	0,00	100,00	4	0,00	0,00	0,00	33,33	66,67
						5	0,00	0,00	0,00	50,00	50,00

The results obtained showed that the ICU professionals are very comfortable with the new system. In general, the questions related with decision making support were scored to 4 or 5 points. This gives a good motivation to continue the work improving the entire system.

6 Conclusion and Future Work

An intelligent scoring system has been presented to support the decisions taken in the ICU environment and, at the same time, improve the patient results. This approach makes it possibility to provide a set of scores calculated / updated in real time. The scoring system proposed processes automatically the scores and adapt the results according to the new values collected, generating new knowledge. The main gains in using this approach can be summarized as:

- The data acquisition, the scores calculation and the results are made in real-time;
- All values are considered - no missing values;
- The data is displayed in a new way – real time charts to compare trends;
- Less human intervention in the scores calculation – less errors;
- The scores are available anywhere and anytime;
- Help decision making process through a continuous scores monitoring - a real-time calculation of the scores according the most recent patient results, allow a quick and better comprehension of the patient condition.

Further work includes the study of other scores (e.g. MEWS) and their impact in the DM models.

Acknowledgments. The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EIA/72819/2006. The work of Filipe Portela was supported by the grant SFRH/BD/70156/2010 from FCT.

References

[1] Keegan, M.T., Gajic, O., Afessa, B.: Severity of illness scoring systems in the intensive care unit. Critical Care Medicine 39, 163 (2011)

[2] Schusterschitz, N., Joannidis, M.: Predictive capacity of severity scoring systems in the ICU. Contributions to Nephrology 156, 92 (2007)

[3] Gago, P., Santos, M.F., Silva, Á., Cortez, P., Neves, J., Gomes, L.: INTCare: a knowledge discovery based intelligent decision support system for intensive care medicine. Journal of Decision Systems (2006)

[4] Strand, K., Flaatten, H.: Severity scoring in the ICU: a review. Acta Anaesthesiologica Scandinavica 52, 467–478 (2008)

[5] Metnitz, P.G.H., Moreno, R.P., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D., Capuzzo, M., Le Gall, J.R.: SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intensive Care Medicine 31, 1336–1344 (2005)

[6] Shabot, M.M.: Automated data acquisition and scoring for JCAHO ICU core measures, p. 674 (2005)

[7] Brenck, F., Hartmann, B., Mogk, M., Junger, A.: Scoring systems for daily assessment in intensive care medicine. Overview, current possibilities and demands on new developments. Der Anaesthesist 57, 189 (2008)

[8] Vincent, J.L., Bruzzi de Carvalho, F.: Severity of illness. Semin Respir. Crit. Care Med. 31, 031–038 (2010)

[9] Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonca, A., Bruining, H., Reinhart, C.K., Suter, P.M., Thijs, L.G.: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine 22, 707–710 (1996)

[10] Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 270, 2957–2963 (1993)

[11] Moreno, R.P., Metnitz, P.G.H., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D., Capuzzo, M., Le Gall, J.R.: SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine 31, 1345–1355 (2005)

[12] Jones, C.: Glasgow coma scale. AJN The American Journal of Nursing 79, 1551 (1979)

[13] Vilas-Boas, M., Santos, M.F., Portela, F., Silva, Á., Rua, F.: Hourly prediction of organ failure and outcome in intensive care based on data mining techniques. Presented at the 12th International Conference on Enterprise Information Systems, Funchal, Madeira, Portugal (2010)

[14] Portela, F., Gago, P., Santos, M.F., Silva, A., Rua, F., Machado, J., Abelha, A., Neves, J.: Knowledge Discovery for Pervasive and Real-Time Intelligent Decision Support in Intensive Care Medicine. Presented at the KMIS 2011- International Conference on Knowledge Management and Information Sharing, Paris, France (2011)

[15] Santos, M.F., Portela, F., Vilas-Boas, M., Machado, J., Abelha, A., Neves, J.: INTCARE - Multi-agent approach for real-time Intelligent Decision Support in Intensive Medicine. In: 3rd International Conference on Agents and Artificial Intelligence (ICAART), Rome, Italy (2011)

[16] Wooldridge, M.: Intelligent agents. In: Multiagent Systems: a Modern Approach to Distributed Artificial Intelligence, pp. 27–77. MIT Press (1999)

[17] Vassiliadis, P., Simitsis, A., Skiadopoulos, S.: Conceptual modeling for ETL processes, pp. 14–21 (2002)

[18] Hooda, J.S., Dogdu, E., Sunderraman, R.: Health Level-7 compliant clinical patient records system, pp. 259–263 (2004)

[19] Abelha, A., Machado, M., Santos, M., Sollari, A., Rua, F., Paiva, M., Neves, J.: Agency for Archive, Integration and Diffusion of Medical Information. In: Proceeding of AIA (2003)

[20] Santos, M.F., Portela, F., Vilas-Boas, M., Machado, J., Abelha, A., Neves, J.: Information Architecture for Intelligent Decision Support in Intensive Medicine. In: 8th International Conference on Applied Computer and Applied Computational Science (ACACOS 2009), vol. 8, pp. 810–819 (2009)

[21] Portela, F., Santos, M.F., Gago, P., Silva, Á., Rua, F., Abelha, A., Machado, J., Neves, J.: Enabling real-time intelligent decision support in intensive care. In: 25th European Simulation and Modelling Conference- ESM 2011, Guimarães, Portugal, p. 446 (2011)

Data Mining in the Study of the Chronic Obstructive Pulmonary Disease

Maribel Yasmina Santos[1], Jorge Cruz[2], and Artur Teles de Araújo[3]

[1] Information Systems Department, Algoritmi Research Centre, University of Minho, Portugal
maribel@dsi.uminho.pt
[2] Faculty of Medicine, University of Lisbon, Portugal
costacruzjorge@gmail.com
[3] Portuguese Lung Foundation, Portugal
artur@telesdearaujo.com

Abstract. Data Mining algorithms have been used to analyse huge amounts of data and extract useful models or patterns from the analysed data. Those models or patterns can be used to support the decision making process in organizations. In the health domain, and besides the support to the decision process, those algorithms are useful in the analysis and characterization of several diseases. This paper presents the particular case of the use of different Data Mining algorithms to support health care specialists in the analysis and characterization of symptoms and risk factors related with the Chronic Obstructive Pulmonary Disease. This is an airflow limitation that is not fully reversible and that affects up to one quarter of the adults with 40 or more years. For this specific study, data from 1.880 individuals were analysed with decision trees and artificial neural networks in order to identify predictive models for this disease. Clustering was used to identify groups of individuals, with the chronic obstructive pulmonary disease, presenting similar risk factors and symptoms. Furthermore, association rules were used to identify correlations among the risk factors and the symptoms. The results obtained so far are promising as several models confirm the difficulties that are normally associated to the diagnosis of this disease and point to characteristics that must be taken into account in its comprehension.

Keywords: business intelligence, data mining, decision models, chronic obstructive pulmonary disease.

1 Introduction

Organizations collect and store huge amounts of data that are used to support decision making, after an appropriate data analysis process. Independently of the application domain, different data analysis techniques can be used to support organizations in this task. This paper addresses the particular reality of a non-profit organization, the Portuguese Lung Foundation (Fundação Portuguesa do Pulmão - FPP), that carry out several activities to collect, store and analyse data related with several diseases. The obtained results are used to analyse and characterize the current reality and to set up campaigns aiming to improve the citizens' quality of life.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 16–30, 2012.

In particular, this paper is focused on the Chronic Obstructive Pulmonary Disease (COPD) for which Data Mining techniques are used to analyse the collected data. The COPD is an airflow limitation that is not fully reversible and that affects up to one quarter of the adults with 40 or more years [1]. This disease is characterized by some of the following symptoms: chronic cough, sputum production and dyspnoea. It can be confirmed in a clinical exam called spirometry, if the obtained values are 80% below of the Forced Expiratory Volume in 1 second (FEV1) and the ratio FEV1/FVC (Forced Vital Capacity) is lower than 0,7 [2]. The risk factors for COPD usually include: masculine gender, tobacco smoke, exposure to dusts and chemicals, air pollution, asthma, and genetic factors as a rare hereditary deficiency of α1-antitrypsin [2].

This disease can be classified in four stages, according to the degree of severity. The first stage, or Mild COPD, is characterized by a FEV1 value above or equal to 80% and the presence, or not, of chronic cough and sputum production. COPD is not usually detected at this first stage. The second stage, or Moderated COPD, is characterized by a value of the FEV1 between 50% and 79%, shortness of breath during exertion, chronic cough and sputum production. The third stage, or Severe COPD, is characterized by a value of the FEV1 between 30% and 49%, greater shortness of breath, reduced exercise capacity and fatigue. The fourth stage, or Very Severe COPD, is characterized by a value of FEV1 below 30% and the presence of chronic respiratory failure [1, 2].

The analysis of the incidence of COPD is needed in order to provide health specialists with decision support indicators. To accomplish this goal, this paper presents the analysis of a data set made available by the FPP with data collected during 2007. The objective of this work is to use several Data Mining algorithms and verify their usefulness in the study of this disease. Namely, decision trees and artificial neural networks for identifying predictive models; clustering to identify groups of individuals, with the COPD, presenting similar risk factors and symptoms; and, association rules to identify correlations among risk factors and symptoms.

The work presented in this paper is related with the development of a Business Intelligence system for the FPP [3], in which an integrated environment for the collection, storage and analysis of data is made available. To the best of our knowledge, no similar system has been proposed and implemented. This business intelligence system includes a data mart model that enhances the data analysis tasks.

All the results presented in this paper were obtained using the Microsoft SQL Server Business Intelligence Development Studio 2008$^{©}$.

This paper is organized as follows. Section 2 presents some related work and recent studies of this disease with Data Mining techniques. Section 3 summarizes the Data Mining algorithms used in this work. Section 4 gives a brief overview of the available data and the transformations carried out to clean and put the data in the proper format for analysis. Section 5 presents the obtained results and Section 6 concludes with some remarks about the described work and guidelines for future work.

2 Related Work

Several studies have been conducted in order to analyse and verify the impact of COPD in individuals and in the society. According to the World Health Organization (WHO), 65 million people around the world have moderate to severe COPD. More than 3 million people died of COPD in 2005, which correspond to 5% of all deaths. In 2002, COPD was the fifth leading cause of death in the world. The total number of deaths from COPD is projected by more than 30% in the next 10 years, being estimated that COPD will become the third leading cause of death in 2030 [4, 5].

The Portuguese data points to a prevalence of COPD at least of 5,4% of all population [6, 7]. Bárbara et al. [8], using the BOLD methodology, found that in the Lisbon Population, above 40 years old, there is a prevalence of 14,2% of COPD in stage I or higher and 6,9% of stage II or higher. The burden of COPD in Portugal has other important indicators, such as the number of patients admitted in hospitals each year. This number duplicated between 1994 and 2007 and now seems to have a tendency to stabilize. COPD remains the second leading cause of hospital admissions by respiratory diseases [9]. The number of deaths by COPD is of 29/100.000/year, putting COPD as the fifth cause of disease deaths [9].

In the scope of this work, special attention is given to the studies undertaken using advanced data analysis techniques like the ones presented in this paper. Esteban et al. [10] used a classification and regression tree (CART) to predict mortality in patients with stable COPD. For this the authors analysed two independent prospective cohorts: a derivation cohort with 611 recruited patients and a validation cohort with 348 patients. All these patients were followed during 5 years. The CART analysis was used to predict the mortality risk using the following covariates from the derivation cohort: age, FEV1, dyspnoea, physical activity, general health and number of hospital admissions for COPD exacerbations in the previous 2 years. The attribute Age provided the first branch of the obtained decision tree. The highest mortality risk (74%) was verified in patients with more than 75 years, higher levels of dyspnoea and FEV1<50%. The patients with the lowest risk of mortality in 5 years (4%) had less than 55 years, FEV1>35% and one or none recent hospitalisations for COPD exacerbations. In this study the authors were able to show that a decision tree that uses variables commonly gathered by physicians can provide a quick assessment of the severity of the disease, as measured by the risk of mortality in 5 years.

In another study, conducted by Paoletti et al. [11], the authors used exploratory data analysis techniques (like Principal Component Analysis) and clustering methods to study a large data set, in order to assess the presence of hidden structures in data corresponding to the different COPD phenotypes observed in clinical practice. This data set included clinical, functional, and radiological data obtained from 415 patients with COPD enrolled at the Respiratory Unit of the Department of Critical Care of the University of Florence. The enrolled patients had mild to severe airflow limitation representing the broad spectrum of COPD presentations. In order to validate the proposed approach, the authors compared the results obtained using a training data set of 415 patients with lung density data acquired in a test data set of 93 patients who undertook a HRCT (High Resolution Computerized Tomography). The obtained results showed that each patient, although being the individual clinical expression of a wide and continuous spectrum of pathologic changes causing expiratory airflow limitation, could be classified as being affected by predominant airway disease or by predominant parenchyma destructive changes. The evaluation of the cluster membership calculated for each COPD patient was considered, by the authors, as a useful profile indicator and may well impact on understanding the results of pharmacologic trials, on clinician's approach to patient treatment, and on deeper knowledge of COPD natural history.

Clustering was also studied by Weatherall et al. [12] analysing several studies that applied clustering techniques to COPD data. For the authors, these techniques seem particularly suited to the study of diseases that express considerable diversity and as such are ideally placed to address the multidimensional complexity apparent in airways disorders. Further cluster analyses, both population-based and clinic-based, will contribute to a greater understanding of the true patterns of airways disorders.

3 Data Mining

Knowledge Discovery in Databases (KDD) is a complex process concerning the discovery of relationships and other descriptions from data. Data Mining refers to the application algorithms used to extract patterns from data without the additional steps of the KDD process e.g. the incorporation of appropriate prior knowledge and the interpretation of results [13].

Different tasks can be performed in the knowledge discovery process, and several techniques can be applied in the execution of each task. Some of these tasks are *classification*, *clustering*, *association*, *prediction*, *estimation* and *summary*. Data Mining tools offer a wide variety of algorithms to choose from. The performance of each technique depends of the task to be carried out, the quality of the available data and, most important, the objective of the discovery. Some of the Data Mining algorithms include *artificial neural networks*, *decision trees*, *association rules* and *clustering* [14].

Artificial Neural Networks are data models that simulate the structure of the human brain. Like the brain, artificial neural networks learn from a set of inputs and adjust the parameters of the model according to this new knowledge to find patterns in data [15]. Artificial neural networks learn from experience and are useful in detecting unknown relationships between a set of input data and an outcome.

Rule induction can be obtained using Decision Trees [16]. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent the values of the target variable.

Another technique used in Data Mining is Association Rules [17]. Association models are models that examine the extent to which values of one field depend on, or are predicted by, values of another field. Association identifies rules about items that appear together in an event.

Clustering is the process of grouping a set of objects into clusters in such a way that objects having high similarity with each other are placed within a cluster, and they are as dissimilar as possible to objects in other clusters [18].

Clustering, as a data mining technique [19, 20], has been widely used to find groups of individuals with similar behaviour. This paper presents another use of clustering, namely in the epidemiological characterization of COPD in Portugal.

4 Available Data

The data set available for this study was collected by the FPP in initiatives undertaken in Portugal in 2007. These initiatives are open to everyone who wants to participate. In them, the participants are asked to answer a questionnaire that integrates questions related to the symptoms and the risk factors of the COPD. The questionnaire also includes information about the geographical location where patients live (in a qualitative form), and their gender, age, height and weight. The result of the spirometry exam is also recorded. The collected data, 1.880 records, were made available in an Excel file. Table 1 shows the available attributes and the possible values for them.

Table 1. Available attributes for analysis

Attribute	Possible Values
Age	Integer, >0
Gender	{m, f}
Locality	Description (string)
Weight	Integer, >0
Height	Integer, >0
Body Mass Index (BMI)	Decimal
Tobacco	{yes, no, ex-smoker}
Noose / Sneeze	{yes, no}
Lacrimation / Itch	{yes, no}
Dry Cough	{yes, no}
Dry Cough more than 3 months	{yes, no}
Daily Expectoration	{yes, no}
Wheezing	{yes, no}
Wheezing last 12 months	{yes, no}
Wheezing with Flu	{yes, no}
Fatigue	{yes, no}
Shortness of Breath	{yes, no}
Allergies	{yes, no}
Rhinitis	{yes, no}
Asthma	{yes, no}
Daily Asthma Medication	{yes, no}
Asthma Medication in Crises	{yes, no}
Suffers from COPD	{yes, no}
Flu more than twice a year	{yes, no}
Pneumonia	{yes, no, do not know}
Pulmonary Tuberculosis	{yes, no, do not know}
BCG Vaccine	{yes, no, do not know}
Flu Vaccine	{yes, no, do not know}
Pneumonia Vaccine	{yes, no, do not know}
Others Vaccines	{yes, no, do not know}
FEV1	Percentage
FEF 25-75	Percentage

After an extensible analysis of the data, missing data fields and errors were identified. Some of the tasks needed to clean (or to prepare) the data set included: i) the labelling of records without geographical localization; ii) the replacing of *null* values on the height and weight with the mode of each one of these attributes; and, iii) the filling of the *do not know* stamp in all categorical attributes containing *null* values.

After the data cleaning process, a transformation phase took place. It was necessary to add the coordinates (x, y) of the geographical locations to each record. This will allow the use of spatial data mining algorithms, in further analysis, taking into consideration the geographical positions of the patients (the places where they live). Although it is out of the scope of this paper, spatial data mining algorithms were already used and some preliminary results can be found in [3].

In this work, a data-driven approach was followed in order to identify a set of models that could be useful in the decision making process. As part of a whole process, the data analysis tools are fundamental in the identification of relevant information to be included in decision support tools. Working towards the achievement of a decision making model for COPD, Fig. 1 presents the key components of a data-driven decision support model.

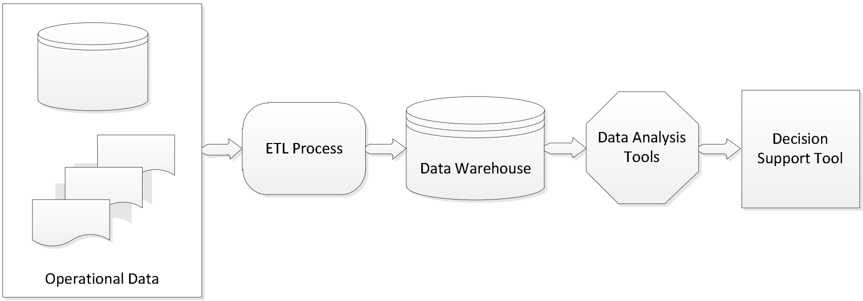


Fig. 1. Data-driven Decision Support Model

This paper emphasises the data analysis tools and the outcomes obtained with Data Mining techniques. In the Data Mining analyses presented in the next section, two different approaches are used. The first one uses all the available data for analysis, providing an overall characterization of the data set. In the second one, only those records (275) related to patients with a FEV1 value lower than 80% are considered. Those patients have a diagnosis of COPD. This allows the characterization of the symptoms and risk factors of individuals with COPD.

For confidentiality reasons, no personal details about the individuals that integrate this data set are provided. Only aggregated results are shown.

5 Obtained Results

All the available data, described previously, was handled in the ETL (Extraction, Transformation and Loading) process in order to load the data to a Data Mart specifically designed to support the FPP in the epidemiological study of COPD. More details about the implemented Business Intelligence System and the supporting Data Mart can be found in [3].

The analysis and characterization of COPD started by the use of decision trees and artificial neural networks to identify models that could be used in a predictive task: whether, or not, an individual can be classified as having COPD based on his/her risk factors, symptoms and other diseases. For this task, several decision trees and artificial neural networks were trained in order to test their confidence in the prediction task. We started by the analysis of the risk factors and their capacity to predict COPD. The attributes used as input were Age Class, Allergies, Bronchial Asthma, Ex-Smoker, Smoker and Gender, and the output attribute was Diagnosis COPD. Different random divisions of the training/test data sets were used, being the obtained results, in terms of confidence of the models, presented in Table 2. The analysis of this table shows that there is no significant difference between the confidence of the models taking into consideration the division of the available data.

Table 2. Confidence of the different models (Risk Factors)

Training Data Set	Test Data Set	Decision Tree	Neural Network
30%	70%	85%	84%
40%	60%	84%	84%
50%	50%	85%	85%
60%	40%	85%	85%
70%	30%	84%	84%

Fig. 2 shows the decision tree obtained for the 60/40 division. In this tree we can see that the decision of being classified as having, or not, COPD is based on the values of the Bronchial Asthma, Ex-Smoker and Age Class attributes. In what concerns the artificial neural network, Fig. 3 presents a discriminant view of the attributes, and corresponding values, used to predict COPD. Analysing this figure, the value with the highest influence in a decision "Favors yes" is having Bronchial Asthma, followed by the Age Class (with individuals with 41 or more years old), being Ex Smoker and having Allergies. Both models, decision tree and artificial neural network, present a similar structure, including in many cases the same attributes in the decision making process.

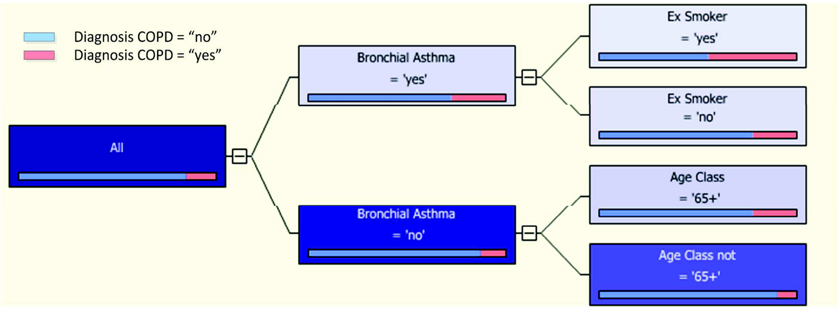
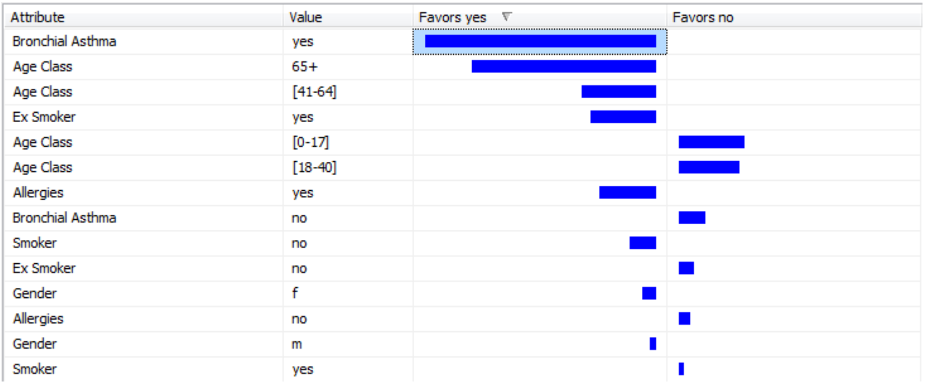


Fig. 2. Decision tree for the diagnosis of COPD based on the risk factors



Attribute	Value	Favors yes ▽	Favors no
Bronchial Asthma	yes	█	
Age Class	65+	█	
Age Class	[41-64]	█	
Ex Smoker	yes	█	
Age Class	[0-17]		█
Age Class	[18-40]		█
Allergies	yes	█	
Bronchial Asthma	no		█
Smoker	no	█	
Ex Smoker	no		█
Gender	f	█	
Allergies	no		█
Gender	m	█	
Smoker	yes		█

Fig. 3. Artificial neural network for the diagnosis of COPD based on the risk factors

Besides the risk factors associated to COPD, there are several symptoms that are usually verified by individuals with COPD. In order to see if it is possible to predict if an individual has, or not, COPD, based on the symptoms, several decision trees and artificial neural networks were trained to verify the obtained models and their prediction capabilities. Starting by the division of the available data set, several combinations were tested. The obtained results show the same behaviour verified for the models obtained for the risk factors. With a division of 30% for training and 70% for testing, both models achieve a confidence of 84%. Using the opposite division, 70% for training and 30% for testing, both models present a confidence of 86%.

Fig. 4 shows the decision tree for this last case. Shortness of Breath and feel More Fatigue Than People of the Same Age are the attributes used by this decision model. By the analysis of the decision tree it is noticeable that there is no obvious symptom, or set of symptoms, which clearly state that an individual suffers from COPD. Looking at the results obtained with the artificial neural network (Fig. 5), these are also the two attributes with more relevance in the decision taken by the obtained model.

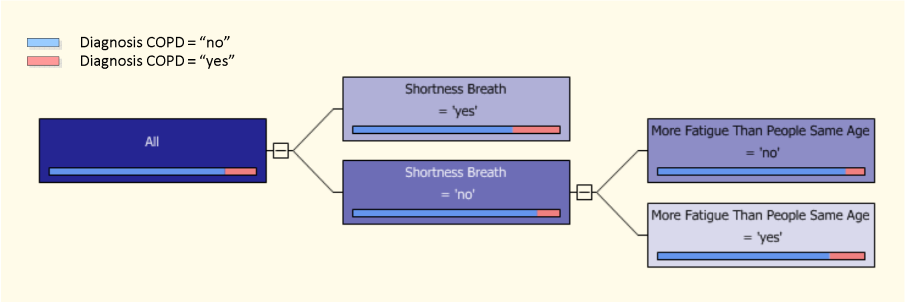
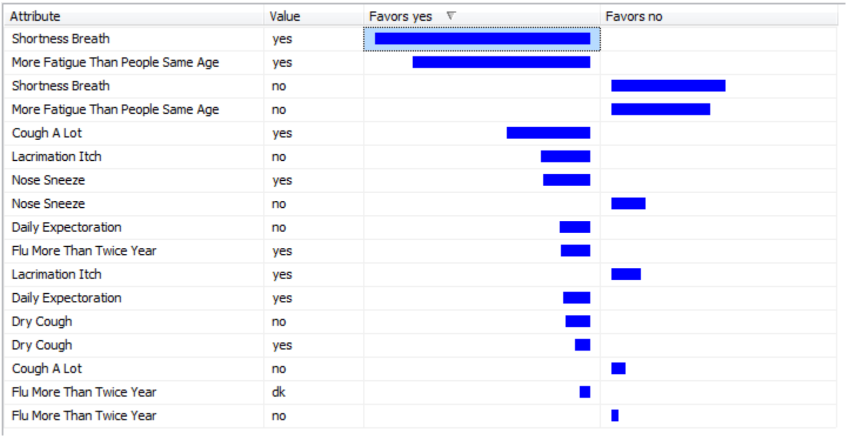


Fig. 4. Decision tree for the diagnosis of COPD based on the symptoms

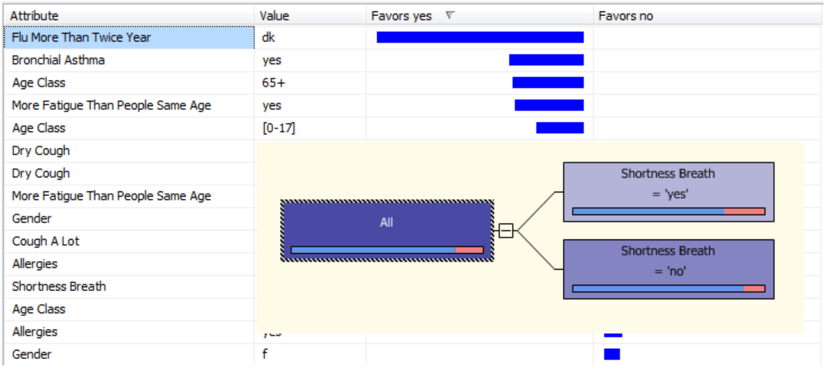


Attribute	Value	Favors yes	Favors no
Shortness Breath	yes	▬	
More Fatigue Than People Same Age	yes	▬	
Shortness Breath	no		▬
More Fatigue Than People Same Age	no		▬
Cough A Lot	yes	▬	
Lacrimation Itch	no	▬	
Nose Sneeze	yes	▬	
Nose Sneeze	no		▬
Daily Expectoration	no	▬	
Flu More Than Twice Year	yes	▬	
Lacrimation Itch	yes		▬
Daily Expectoration	yes	▬	
Dry Cough	no	▬	
Dry Cough	yes	▬	
Cough A Lot	no		▬
Flu More Than Twice Year	dk	▬	
Flu More Than Twice Year	no		▬

Fig. 5. Artificial neural network for the diagnosis of COPD based on the symptoms

The models obtained so far seem to present some limitations in their prediction capabilities if we take into consideration that none of them was able to clearly predict the COPD cases (Diagnosis COPD =”yes”). These results can be influenced by the fact that we are dealing with a non homogeneous data set in terms of the number of cases with “yes” and “no” in the COPD diagnosis. The data set includes 275 positive cases of COPD in a total of 1.880 cases. A balancing of the data set was tested, by randomly reducing the number of records with negative cases, but the obtained results showed no improvements, as the decision models were the same.

After the analysis of risk factors and symptoms, separately, we tried the identification of models that combine both in the prediction of COPD. Although the algorithms have more attributes to analyse, the obtained results decreased the confidence of the decision tree and the artificial neural network to 84%. This value is not far from the obtained previously, but while the decision tree takes the Shortness of Breath as the main attribute for the decision, the artificial neural network considers Flu More Than Twice a Year (Fig. 6).



Attribute	Value	Favors yes	Favors no
Flu More Than Twice Year	dk	▬	
Bronchial Asthma	yes	▬	
Age Class	65+	▬	
More Fatigue Than People Same Age	yes	▬	
Age Class	[0-17]	▬	
Dry Cough			
Dry Cough			
More Fatigue Than People Same Age			
Gender			
Cough A Lot			
Allergies			
Shortness Breath			
Age Class			
Allergies	yes		▬
Gender	f		▬

Fig. 6. Analysing risk factors and symptoms

In order to improve the obtained models it will be necessary to collect more data, mainly associated to individuals with COPD. Besides this data collection task, which is out of the scope of this paper, we proceed with the presentation of other relevant models that were obtained and that can be used to better understand and treat people with COPD.

In terms of clustering, risk factors and symptoms were also explored in order to identify groups of individuals that, having COPD, present similar characteristics. Starting by the analysis of risk factors, in order to understand how they are related in the patients that present COPD, Fig. 7 shows the 8 obtained clusters. Cluster 1 is characterized by integrating individuals of several Ages that do not present Allergies, do not have Bronchial Asthma, never Smoked and are mostly Female. In this cluster we have 53 individuals that suffer from COPD but that do not have any apparent risk factors of this disease. Cluster 3, with 37 individuals with COPD, integrates mainly persons with an Age between 41 and 64 years old, all suffering from Allergies and most of them also suffering from Bronchial Asthma. The 8 obtained clusters illustrate very different profiles in the characterization of individuals with COPD.

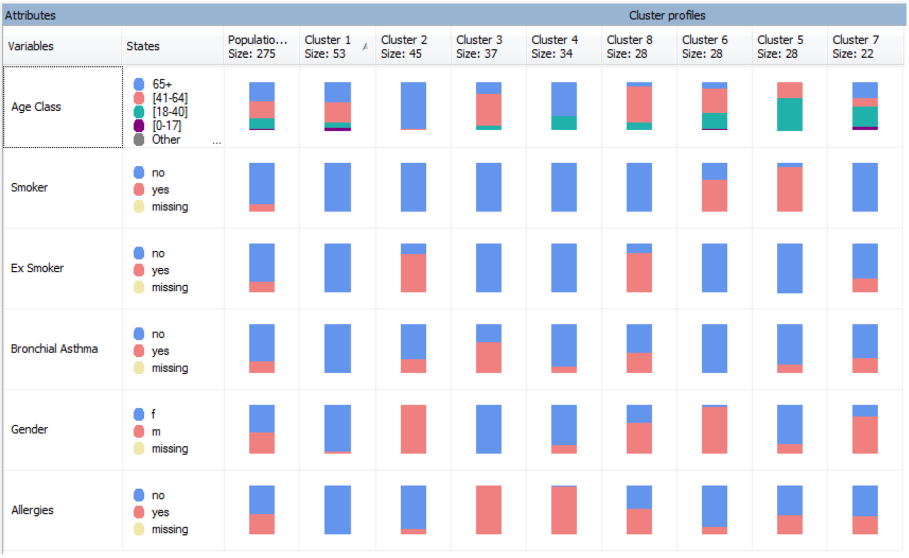
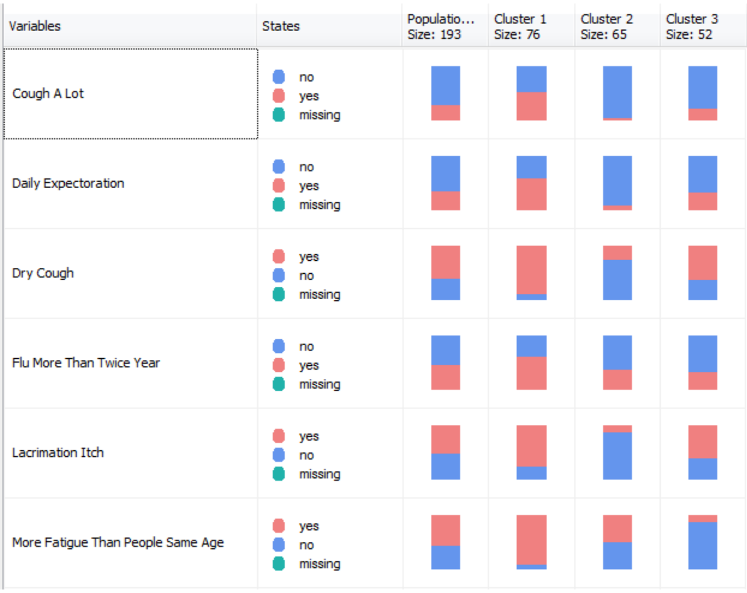


Fig. 7. Clustering results for risk factors

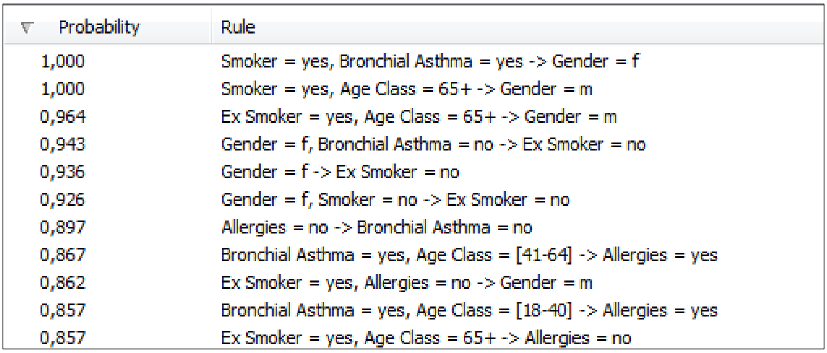
Doing a clustering analysis considering some of the symptoms that are usually associated with COPD, we found the 3 clusters that are shown in Fig. 8. Cluster 1 integrates individuals that Cough a lot, have Daily expectoration, present Dry cough, have Flu more than twice a year and have Lacrimation and Itch. Cluster 2 includes individuals that in the majority of the cases do not present any of these symptoms. Cluster 3 presents a combination of these symptoms.



Variables	States	Populatio... Size: 193	Cluster 1 Size: 76	Cluster 2 Size: 65	Cluster 3 Size: 52
Cough A Lot	● no ● yes ● missing				
Daily Expectoration	● no ● yes ● missing				
Dry Cough	● yes ● no ● missing				
Flu More Than Twice Year	● no ● yes ● missing				
Lacrimation Itch	● yes ● no ● missing				
More Fatigue Than People Same Age	● yes ● no ● missing				

Fig. 8. Clustering results for symptoms

Another analysis undertaken in the study of the risk factors and symptoms of COPD was to verify, using an association rule algorithm, if it is possible to correlate risk factors and symptoms. Starting by the risk factors, it was possible to obtain a total of 28 rules with a probability between 75% and 100%. From this set, we point out the 11 association rules presented in Fig. 9. Looking at the first rule, we can say that if the individual is a Smoker and has Bronchial Asthma then it is a Female with a confidence of 100%. In case it is also a Smoker with 65 or more years old than it is a Male with a probability of 100%. All these rules are associated to individuals with COPD.



▽ Probability	Rule
1,000	Smoker = yes, Bronchial Asthma = yes -> Gender = f
1,000	Smoker = yes, Age Class = 65+ -> Gender = m
0,964	Ex Smoker = yes, Age Class = 65+ -> Gender = m
0,943	Gender = f, Bronchial Asthma = no -> Ex Smoker = no
0,936	Gender = f -> Ex Smoker = no
0,926	Gender = f, Smoker = no -> Ex Smoker = no
0,897	Allergies = no -> Bronchial Asthma = no
0,867	Bronchial Asthma = yes, Age Class = [41-64] -> Allergies = yes
0,862	Ex Smoker = yes, Allergies = no -> Gender = m
0,857	Bronchial Asthma = yes, Age Class = [18-40] -> Allergies = yes
0,857	Ex Smoker = yes, Age Class = 65+ -> Allergies = no

Fig. 9. A subset of the identified association rules for risk factors

Analysing the strongest links between the several attributes and their corresponding values, Fig. 10 shows the dependencies among them. As we can see, the fact of being a Smoker can be used to predict an Age Class of [18-40] or [41-64].

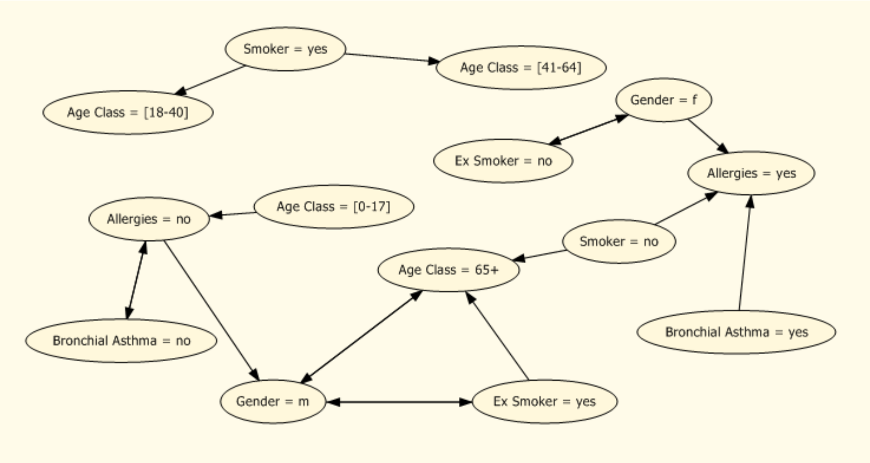
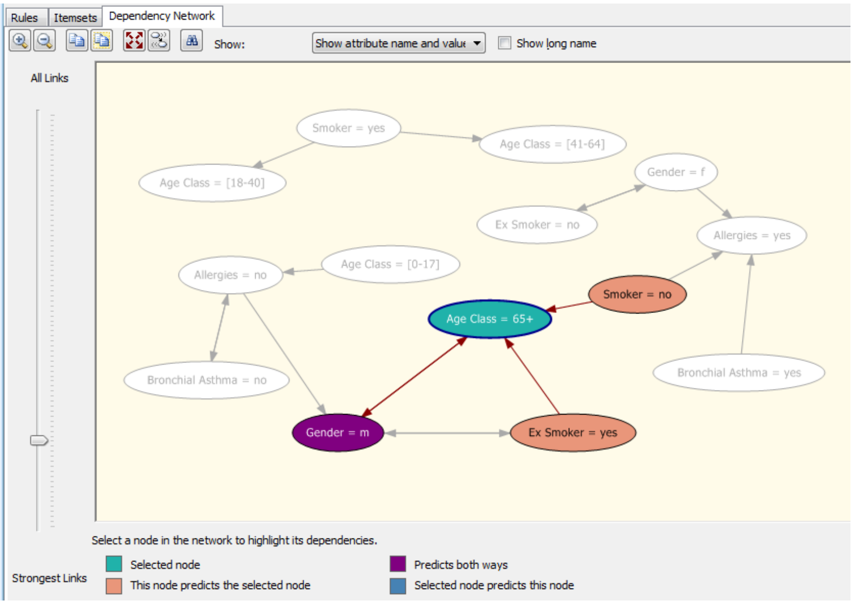


Fig. 10. Strongest links between the several risk factors

In order to facilitate the comprehension of these relationships (Fig. 10), the user can select an item of the dependency network and better analyse the obtained results. Fig. 11 shows one example selecting the Age Class = 65+ node.



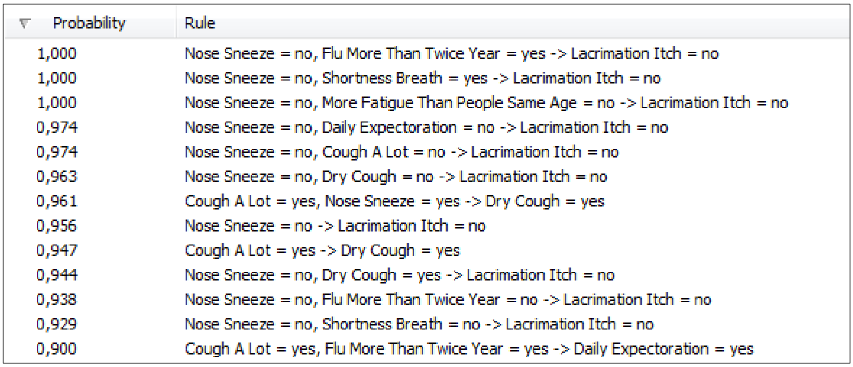
Show: Show attribute name and value □ Show long name

Select a node in the network to highlight its dependencies.

🟩 Selected node
🟧 This node predicts the selected node
🟪 Predicts both ways
🟦 Selected node predicts this node

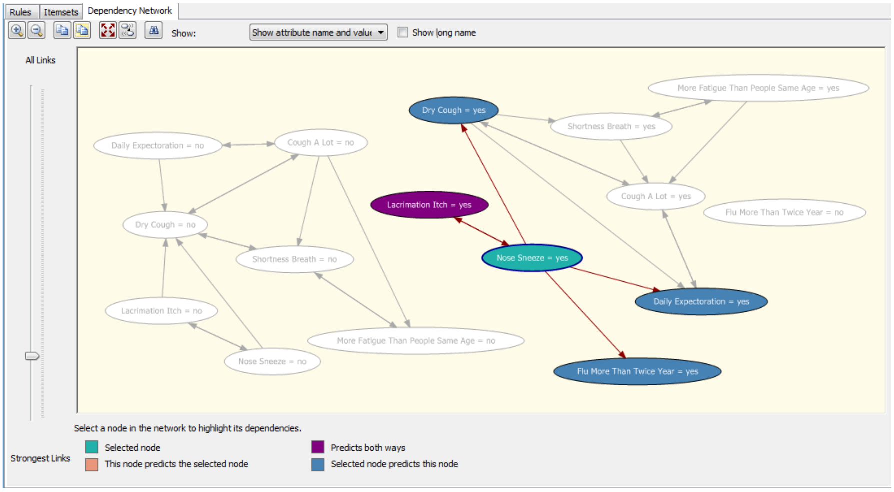
Fig. 11. An example of one strong link (risk factors)

In terms of symptoms, the identification of the association rules, existing among them, was also carried out. The obtained results allowed the identification of 35 rules with a probability between 75% and 100%. Limiting this number of rules to those that have a probability higher or equal to 90%, 13 rules are obtained (Fig. 12). Looking at the strongest links among the values of the attributes, Fig. 13 depicts the particular case of the Nose Sneeze = yes node, being predicted by the Lacrimation Itch = yes node and being responsible for the prediction of the Dry Cough = yes, Daily Expectoration = yes and Flu More Than Twice Year = yes.



Probability	Rule
1,000	Nose Sneeze = no, Flu More Than Twice Year = yes -> Lacrimation Itch = no
1,000	Nose Sneeze = no, Shortness Breath = yes -> Lacrimation Itch = no
1,000	Nose Sneeze = no, More Fatigue Than People Same Age = no -> Lacrimation Itch = no
0,974	Nose Sneeze = no, Daily Expectoration = no -> Lacrimation Itch = no
0,974	Nose Sneeze = no, Cough A Lot = no -> Lacrimation Itch = no
0,963	Nose Sneeze = no, Dry Cough = no -> Lacrimation Itch = no
0,961	Cough A Lot = yes, Nose Sneeze = yes -> Dry Cough = yes
0,956	Nose Sneeze = no -> Lacrimation Itch = no
0,947	Cough A Lot = yes -> Dry Cough = yes
0,944	Nose Sneeze = no, Dry Cough = yes -> Lacrimation Itch = no
0,938	Nose Sneeze = no, Flu More Than Twice Year = no -> Lacrimation Itch = no
0,929	Nose Sneeze = no, Shortness Breath = no -> Lacrimation Itch = no
0,900	Cough A Lot = yes, Flu More Than Twice Year = yes -> Daily Expectoration = yes

Fig. 12. A subset of the identified association rules for symptoms



Select a node in the network to highlight its dependencies.

🟦 Selected node
🟧 This node predicts the selected node
🟪 Predicts both ways
🟦 Selected node predicts this node

Fig. 13. An example of one strong link (symptoms)

After the use of different data analysis techniques, several relevant findings were obtained. In order to make them useful in the decision making process, proper decision tools need to be developed and made available to the health care specialists. This step will complete the decision model previously presented in Fig. 1.

6 Conclusions and Future Work

This paper presented the use of several data mining techniques for the analysis of data collected by the Portuguese Lung Foundation (FPP – *Fundação Portuguesa do Pulmão)*. The objectives set to this work were to analyse and characterize the COPD (Chronic Obstructive Pulmonary Disease) using decision trees, artificial neural networks, clustering and association rules.

Data from 1.880 patients were available from the FPP's data set. Each patient answered a questionnaire and made a clinical exam called spirometry. The several analyses reinforced the knowledge that this disease is very difficult to diagnose without the spirometry exam. That happens because despite we have confirmed some risk factors related to COPD and identified some patterns for this disease, there seems to be a kind of contradiction with their symptoms. For example, many patients do not have any symptom associated to fatigue and cough, which are typical symptoms of this respiratory disease. Therefore, the difficulty in diagnosing the disease also reinforce the conclusion that a good prevention of COPD, making a spirometric exam periodically, is essential for people who fall within the risk factors of this disease.

The used data mining techniques showed to be useful in the analysis of the available data, although some of the obtained models can be improved with the analysis of a more balanced data set, in terms of the positive and negative cases of COPD.

For future work, and besides the collection and integration of more data in the data mining analyses, we intend to incorporate data about other diseases, like pneumonia, and analyse the possible relationships between the risk factors and symptoms of these two diseases.

This study may be considered as an example showing possible applications of data mining techniques to investigate clinical aspects of chronic pathologies where a decision model is usually missing. In this work, a data-driven approach was followed in order to avoid any bias that could be introduced by previous knowledge of the domain application. This approach could be very useful in the definition of the patients' profile.

Acknowledgements. This work was funded by FEDER *through the Programa Operacional Fatores de Competitividade – COMPETE* and by national funding through FCT (*Fundação para a Ciência e Tecnologia*), project: FCOMP-01-0124-FEDER-022674.

References

1. GOLD, Global Strategy for Diagnosis, Management, and Prevention of COPD. Technical Report, Global Initiative for Chronic Obstructive Lung Disease (2010)
2. Pauwels, R.A., Buist, A.S., Calverley, P.M.A., Jenkins, C.R., Hurd, S.S.: Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine 163, 1256–1276 (2001)

3. Dinis, R., Ribeiro, A., Santos, M.Y., Cruz, J., Araújo, A.T.D.: A Business Intelligence Infrastructure Supporting Respiratory Health Analysis. In: Proceedings of the First International Conference on Business Intelligence and Technology (BusTECH 2011), Rome, Italy (2011)
4. Buist, A.S., McBurnie, M.A., et al.: International variation in the prevalence of COPD (the BOLD study): a population-based prevalence study. Lancet 370(9589), 741–750 (2007)
5. Lamprecht, B., McBurnie, M.A., et al.: COPD in never smokers: results from the population-based burden of obstructive lung disease study. Chest 139(4), 752–763 (2011)
6. Cardoso, J., et al.: Prevalence of Chronic Obstructive Pulmonary Disease in Portugal. Am J. Resp. Crit. Care Med. 167(7), 23 (2003)
7. Borges, M., Gouveia, M., et al.: Carga de Doença atribuível ao tabagismo em Portugal. Rev. Port. Pneumol. 15(6), 951–1004 (2009) (in Portuguese)
8. Bárbara, C., Rodrigues, F., et al.: COPD Prevalence in Portugal. The Burden of Obstructive Lung Disease Study. In: European Respiratory Society Annual Congress, Barcelona (2010)
9. RONDR, Relatório do Observatório Nacional das Doenças Respiratórias. Portuguese Lung Foundation. pp. 142–149 (2010) (in Portuguese)
10. Esteban, C., Arostegui, I., et al.: Development of a decision tree to assess the severity and prognosis of stable COPD. European Respiratory Journal 38(6), 1294–1300 (2011)
11. Paoletti, M., Camiciottoli, G., et al.: Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes. Journal of Biomedical Informatics 42, 1013–1021 (2009)
12. Weatherall, M., Shirtcliffe, P., Travers, J., Beasley, R.: Use of cluster analysis to define COPD phenotypes. European Respiratory Journal 36(3), 472–474 (2010)
13. Fayyad, U., Uthurusamy, R.: Data Mining and Knowledge Discovery in Databases. Communications of the ACM 39(11), 24–26 (1996)
14. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2001)
15. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U.M., et al. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 1–34. The MIT Press, Massachusetts (1996)
16. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
17. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD Conference on Management of data, Washington, DC (1993)
18. Zaït, M., Messatfa, H.: A comparative study of clustering methods. Future Generation Computer Systems 13(2), 149–159 (1997)
19. Cios, K., Pedrycz, W., Swiniarski, R.: Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers (1998)
20. Groth, R.: Data Mining: Building Competitive Advantage. Prentice Hall PTR (2000)

Practical Problems and Solutions in Hospital Information System Data Mining

Miroslav Bursa[1], Lenka Lhotska[1], Vaclav Chudacek[1], Jiri Spilka[1], Petr Janku[2], and Martin Huser[2]

[1] Department of Cybernetics,
Faculty of Electrical Engineering,
Czech Technical University in Prague, Czech Republic
[2] Obstetrics and Gynaecology Clinic,
University Hospital in Brno, Czech Republic

Abstract. Information mining from textual data becomes a very challenging task when the structure of the text record is very loose without any rules. Doctors often use natural language in medical records. Therefore it contains many ambiguities due to non-standard abbreviations and synonyms. The medical environment itself is also very specific: the natural language used in textual description varies with the personality creating the record (there are many personalized approaches), however it is restricted by terminology (i.e. medical terms, medical standards, etc.). Moreover, the typical patient record is filled with typographical errors, duplicates, ambiguities, syntax errors and many nonstandard abbreviations.

This paper describes the process of mining information from loosely structured medical textual records with no apriori knowledge. The paper concerns mining a large dataset of ∼50,000–140,000 records × 20 attributes in relational database tables, originating from the hospital information system (thanks go to the University Hospital in Brno, Czech Republic) recording over 11 years. This paper concerns only textual attributes with free text input, that means 650,000 text fields in 16 attributes. Each attribute item contains approximately 800–1,500 characters (diagnoses, medications, anamneses, etc.). The output of this task is a set of ordered/nominal attributes suitable for automated processing that can help in asphyxia prediction during delivery.

The proposed technique has an important impact on reduction of the processing time of loosely structured textual records for experts.

Note that this project is an ongoing process (and research) and new data are still received from the medical facility, justifying the need for robust and fool-proof algorithms.

In the preliminary analysis of the data, classical approaches such as basic statistic measures, word (and word sequence) frequency analysis, etc., have been used to simplify the textual data and provide a preliminary overview of the data. Finally, an ant-inspired self-organizing approach has been used to automatically provide a simplified dominant structure, presenting structure of the records in the human readable form that can be further utilized in the mining process as it describes the vast majority of the records.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 31–39, 2012.

Keywords: Swarm Intelligence, Ant Colony, Text Data Mining, Information Retrieval, Hospital Information System.

1 Introduction

1.1 Motivation

In many industrial, business, healthcare and scientific areas we witness the boom of computers, computational appliances, personalized electronics, high-speed networks, increasing storage capacity and data warehouses. Therefore a huge amount of various data is transferred and stored, often mixed from different sources, containing different data types, unusual coding schemes, and seldom come without any errors (or noise) and omissions. Massively parallel distributed storage systems are used nowadays to provide computational nodes with data in reasonable time.

There are also problems with on-time data availability for a computational node. Especially in text processing, the impact of automated methods is crucial. In contrary to classical methods, nature-inspired methods offer many techniques, that can increase speed and robustness of classical methods.

1.2 Nature Inspired Methods

Nature inspired metaheuristics play an important role in the domain of artificial intelligence, offering fast and robust solutions in many fields (graph algorithms, feature selection, optimization, clustering, feature selection, etc). Stochastic nature inspired metaheuristics have interesting properties that make them suitable to be used in data mining, data clustering and other application areas.

In the last two decades, many advances in the computer sciences have been based on the observation and emulation of processes of the natural world. The origins of *bioinspired informatics* can be traced to the development of perceptrons and artificial life, which tried to reproduce the mental processes of the brain and biogenesis respectively, in a computer environment [1]. Bioinspired informatics also focuses on observing how the nature solves situations that are similar to engineering problems we face.

With the boom of high-speed networks and increasing storage capacity of database clusters and data warehouses, a huge amount of various data can be stored. *Knowledge discovery* and *Data mining* is not only an important scientific branch, but also an important tool in industry, business and healthcare. These techniques target the problematic of processing huge datasets in reasonable time – a task that is too complex for a human. Therefore computer-aided methods are investigated, optimized and applied, leading to the simplification of the processing of the data. The main goal of computer usage is data reduction preserving the statistical structure (clustering, feature selection), data analysis, classification, data evaluation and transformation.

Ant Algorithms. Ant colonies inspired many researchers to develop a new branch of stochastic algorithms: *ant colony inspired algorithms*. Based on the ant metaphor, algorithms for both static and dynamic combinatorial optimization, continuous optimization and clustering have been proposed. They show many properties similar to the natural ant colonies, however, their advantage lies in incorporating the mechanisms, that allowed the whole colonies to effectively survive during the evolutionary process.

1.3 Knowledge Extraction

Several techniques to extract knowledge from raw data have been developed in the past. These techniques have various and multiple origins: some result from the statistical analysis of the data, the regressions, decision trees, etc.; some resulting from the artificial intelligence such as the expert systems, intelligent agents, fuzzy logic, etc.

Plenty of nature inspired methods are studied and developed in present. One category is represented by methods, that are inspired by the behavior of ant colonies. These methods have been applied to many problems (often NP-hard). Review can be seen in [6] and [2]. We concentrate on the state-of-the-art nature methods inspired by the social behavior of insect communities, by the swarm intelligence, brain processes and other real nature processes.

Text Extraction. The accuracy for relation extraction in journal text is typically about 60 % [7]. A perfect accuracy in text mining is nearly impossible due to errors and duplications in the source text. Even when linguists are hired to label text for an automated extractor, the inter-linguist disparity is about 30 %. The best results are obtained via an automated processing supervised by a human [9].

Onthologies have become an important means for structuring knowledge and building knowledge-intensive systems. For this purpose, efforts have been made to facilitate the ontology engineering process, in particular the acquisition of onthologies from texts.

Additionally, medical records are specific. Although the medical doctors are members of an association, no real measures relating the unification of terminology and disambiguation is really made. The level of semantic interoperability is low, making the automated information retrieval a really nontrivial task.

Doctors often use natural language in medical records. Therefore it contains many ambiguities due to non-standard abbreviations and synonyms. For example, the information of *diabetes mellitus* is often expressed as *DM*, *DM2*, *DM 2*, *dia II*, *DMII*, etc. Even the national nomenclature (a translation of the ICD10 system) contains nonstandard and ambiguous abbreviations. The medical environment itself is also very specific: the natural language used in textual description varies with the personality creating the record (there are many personalized approaches), however it is (not strictly) restricted by terminology (i.e. medical terms, medical standards, etc.).

2 Input Dataset Overview

The dataset consists of a set of approx. 50 to 120 thousand records (structured in different relational DB tables; some of them are not input, therefore the range is mentioned) × approx. 20 attributes. Each record in an attribute contains about 800 to 1500 characters of text (diagnoses, patient state, anamneses, medications, notes, references to medical stuff, etc.). For textual mining, 16 attributes are suitable (contain sufficiently large corpus).

The database export is a 10+ year export from the hospital information system. Anonymization has been performed in order not to reveal the sensitive patient data. As the patient ID is considered sensitive data, the records and other signals available have been referenced using a MD5 hash.

The overview of one small (in field length) attribute is visualized in Fig. [1]. Only a subsample (about 5 %) of the dataset could be displayed in this paper, as the whole set would render into a incomprehensible black stain. The vertices (literals) are represented as colored circle, the size reflects the literal (i.e. word) frequency. Edges represent transition states between literals (i.e. the sequence of 2 subsequent words in a sentence/record); edge stroke shows the transition rate (probability) of the edge. The same holds for all figures showing the transition graph, only a different visualization approach has been used.

It is clear, that human interpretation and analysis of the textual data is very fatiguing, therefore any computer aid is highly welcome.

3 Motivation

The task of this work is to provide the researchers with a quick automated or semi-automated view on the textual records. Textual data are not easy to visualize. The word frequency method is simple, but did not provide easily interpretable data. A frequency of multiple words is also a valuable input, however contains many duplications and does not really contribute in the process of definition of regular expressions for further mining. Therefore we decided to extract information in the form of a transition graph.

Such graphs allow as to induce a set of rules for information retrieval. These rules serve for extraction of (boolean/nominal) attributes from the textual rules. These attributes are used in automated rule discovery and can be further used for recommendation. The overall goal of the project is asphyxia prediction during delivery. High asphyxia might lead to several brain damage of the neonate and when predicted, caesarean section might be indicated on time.

4 Graph Explanation

In this paper we describe *transition graphs*. These are created for each attribute. An attribute consists of many records in form of a sentence. By *sentence* we hereby mean a sequence of literals, not a sentence in a linguistic form. The records are compressed – unnecessary words (such as verbs *is*, *are*) are omitted.

In this paper, only the attribute describing the anesthetics during deliveries visualized, as it is the simplest one.

Vertices of the transition graph represent the words (separated by spaces) in the records. For each word (single or multiple occurrence) a vertex is created and its potence (number of occurrences is noted). For example, the words *mesocaine*, *anesthetics*, *not*, *mL* form a vertex. Note that also words as *mesocain*, *mezokain* and other versions of the word *mesocaine* are present. For a number (i.e. sequence of digits) a special literal *_NUMBER_* is used.

Edges are created from single records (sentences entered). For example the sentence *mesocaine 10 mL* would add edges from vertex *mesocaine* to vertex *_NUMBER_* and from vertex *_NUMBER_* to the vertex *mL* (or the edge count is increased in case it exists). For all records, the count of the edges is also useful. It provides an overview on the inherent structure of the data – the most often word transitions.

Note that only a small subsection of the records of only one attribute is visualized in the paper. When displaying all records, the graphs are unprintable in the common paper formats and usually render as a black stain, therefore the ink-to-information ratio is very high. But it is totally unreadable. Images are supplied in a vector format, so the should be zoomed in correctly.

5 Nature Inspired Techniques

Social insects, i. e. ant colonies, show many interesting behavioral aspects, such as self-organization, chain formation, brood sorting, dynamic and combinatorial optimization, etc. The coordination of an ant colony is of local nature, composed mainly of indirect communication through pheromone (also known as *stigmergy*, the term has been introduced by Grassé et al. [8]), although direct interaction communication from ant to ant (in the form of antennation) and direct communication have also been observed [11].

The high number of individuals and the decentralized approach to task coordination in the studied species means that ant colonies show a high degree of parallelism, self-organization and fault tolerance. In studying these paradigms, we have high chance to discover inspiration concepts for many successful metaheuristics.

5.1 Ant Colony Optimization

Ant Colony Optimization (ACO) [6] is an optimization technique that is inspired by the foraging behavior of real ant colonies. Originally, the method was introduced for the application to discrete and combinatorial problems.

Ant Colony Methods for Clustering. Several species of ant workers have been reported to form piles of corpses (cemeteries) to clean up their nests. This aggregation phenomenon is caused by attraction between dead items mediated by the ant workers.

This approach has been modeled in the work of Deneubourg et al. [5] and in the work of Lumer and Faieta [10] to perform a clustering of data.

ACO_DTree Method. The ACO_DTree method is a hybrid evolutionary approach for binary decision tree construction [4]. The tree is induced using the known data and can be further used for unsupervised clustering later: each leaf of the classification tree can be interpreted as a cluster. The algorithm uses a population of classification trees that is evolved using an evolutionary approach. Creation of the trees is driven by a pheromone matrix, which uses the ACO paradigm.

This approach has been utilized (with improvements and adaptation to the specific problem area) to simplify the structure of the vast dataset by finding the most important state transitions between literals, producing a probabilistic transitional model. The output structure is presented to the analyst for further processing/iteration.

For clustering, the ACO_DTree method [4,3] and ACO inspired clustering [10] variations have been successfully used. A self-organizing map has also been tested, but performed poorly.

New solutions are constructed continuously. It is a stochastic decision process based on the pheromone amount sensed by the ants. As in nature, the pheromone slowly evaporates over time (over iterations) in order to avoid getting stuck in local minimum and to adapt to dynamically changing environment. Daemon actions represent *background* actions which consist mainly of pheromone deposition. The amount is proportional to the quality of solution (and appropriate adaptive steps).

Main parameters of the algorithm are (with major importance to the method proposed): pheromone lay rate, pheromone evaporate rate, number of solutions created (number of ants), number of iterations, etc.

6 Automated Processing

Automated layout of transition graph is very comfortable for an expert, however the contents of the attribute is so complicated, that a human intervention is inevitable. Examples of automated layout can be seen in Fig. [1].

The figure Fig. [1] shows a transitional graph where only positioning based on the word distance from the sentence start is used. Although it migh look correct, note that the same words are mispositioned in the horizontal axis.

7 Expert Intervention

A human intervention and supervision over the whole project is indiscutable. Therefore also human (expert) visualization of the transition graph has been studied.

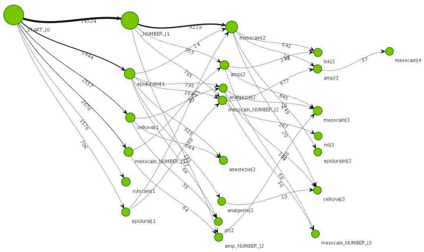


Fig. 1. A fully automated transition graph showing the most important relations in one textual attribute. No clustering has been used. The layout is based on the word distance from the start of the sentence. Note the mis-alignment of the similar/same words. Refer to section [2].

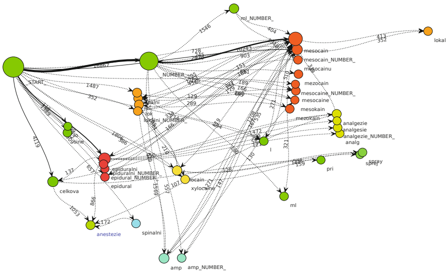


Fig. 2. An expert (human) organized transition graph (sub-graph) showing the most important relations in one textual attribute. Refer to section [2].

The vertices in a human-only organization are (usually) organized depending on the position in the text (distance from the starting point) as the have the highest potence. Number literal (a wildcard) had the highest potence, as many quantitative measures are contained in the data (age, medication amount, etc.). Therefore it has been fixed to the following literal, spreading into the graph via multiple nodes (i.e. a sequence *mesocain 10 mL* become two vertices – *mesocain_NUMBER_* and *mL*). This allowed to organize the chart visualization in more logical manner. Time needed to organize such graph was about 5–10 minutes. The problem is that the transition graph contains loops, therefore the manual organization is not straightforward.

An aid of a human expert has been used in semi-automated approach (see Fig. [3] where the automated layout has been corrected by the expert. The correction time has been about 20–30 seconds only.

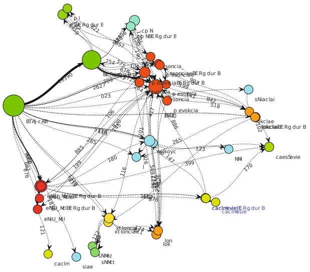


Fig. 3. A semi-automated (corrected by a human expert) organized transition graph showing the most important relations in one textual attribute. Refer to section [2].

8 Results and Conclusion

The main advantage of the nature inspired concepts lies in automatic finding relevant literals and group of literals that can be adopted by the human analysts and furthermore improved and stated more precisely. The use of induced probabilistic models in such methods increased the speed of loosely structured textual attributes analysis and allowed the human analysts to develop lexical analysis grammar more efficiently in comparison to classical methods. The speedup (from about 5–10 minutes to approx 20–30 seconds) allowed to perform more iterations, increasing the yield of information from data that would be further processed in rule discovery process. However, the expert intervention in minor correction is still inevitable. The results of the work are adopted for rule discovery and are designed to be used in expert recommendation system. A secondary output of this project is the gained knowledge for design of interoperable medical systems.

9 Discussion and Future Work

The future work is to evaluate the DB analyst's utilization and aid of such graphs in more accurate way. The graphs serve as a bases for extraction rule proposal. However the only relevant measure is the time to reorganize the transitional graphs. The subjective opinion is very expressive and is not coherent. Next, the semantic meaning of the attributes will be extracted and verified followed by rule discovery mining.

Acknowledgements. This research project has been supported by the research programs number MSM 6840770012 "Transdisciplinary Research in the Area of Biomedical Engineering II" of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of the Czech Republic and by the project number NT11124-6/2010 "Cardiotocography evaluation by means of artificial intelligence" of the Ministry of Health Care. This work has been developed in the BioDat research group `http://bio.felk.cvut.cz`.

References

1. Adami, C.: Introduction to Artificial Life. Springer (1998)
2. Blum, C.: Ant colony optimization: Introduction and recent trends. Physics of Life Reviews 2(4), 353–373 (2005)
3. Bursa, M., Huptych, M., Lhotska, L.: Ant colony inspired metaheuristics in biological signal processing: Hybrid ant colony and evolutionary approach. In: Biosignals 2008-II, vol. 2, pp. 90–95. INSTICC Press, Setubal (2008)
4. Bursa, M., Lhotska, L., Macas, M.: Hybridized swarm metaheuristics for evolutionary random forest generation. In: Proceedings of the 7th International Conference on Hybrid Intelligent Systems 2007 (IEEE CSP), pp. 150–155 (2007)
5. Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., Chretien, L.: The dynamics of collective sorting robot-like ants and ant-like robots. In: Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats, pp. 356–363. MIT Press, Cambridge (1990)
6. Dorigo, M., Stutzle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
7. Freitag, D., McCallum, A.K.: Information extraction with hmms and shrinkage. In: Proceedings of the AAAI Workshop on Machine Learining for Information Extraction (1999)
8. Grasse, P.P.: La reconstruction du nid et les coordinations inter-individuelles chez bellicositermes natalensis et cubitermes sp. la thorie de la stigmergie: Essai d'interprtation des termites constructeurs. Insectes Sociaux 6, 41–81 (1959)
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the ICML, pp. 282–289 (2001); text processing: interobserver agreement among linquists at 70
10. Lumer, E.D., Faieta, B.: Diversity and adaptation in populations of clustering ants. In: From Animals to Animats: Proceedings of the 3th International Conference on the Simulation of Adaptive Behaviour, vol. 3, pp. 501–508 (1994)
11. Trianni, V., Labella, T.H., Dorigo, M.: Evolution of Direct Communication for a *Swarm-bot* Performing Hole Avoidance. In: Dorigo, M., Birattari, M., Blum, C., Gambardella, L.M., Mondada, F., Stützle, T. (eds.) ANTS 2004. LNCS, vol. 3172, pp. 130–141. Springer, Heidelberg (2004)

Using Generic Meta-Data-Models for Clustering Medical Data

Dominic Girardi[1], Michael Giretzlehner[1], and Josef Küng[2]

[1] RISC Software GmbH - Research Unit Medical Informatics, Hagenberg
firstname.lastname@risc.uni-linz.ac.at
[2] Institute for Application Oriented Knowledge Processing, JKU Linz
jkueng@faw.uni-linz.ac.at

Abstract. We present a generic, meta-model based data storage system for research, clinical studies or disease registers, which is enabled to store data of almost arbitrary structure. The system is highly costumizeable and allows the user to set up a professional web-based data acquisition system including administration area, data input forms, overview tables and statistics within hours. Furthermore, we evaluated a number of clustering algorithms regarding their ability to cluster the stored datasets for similarity search and further statistical analysis.

1 Introduction

An average university hospital produces 6 million documents a year [1]. An effort, that hardly seems arguable. Nevertheless, this amout of data is necessary to provide a complete history of all treatments of patients, which is a basic requirement for an accurate treatment. Despite or even because of its huge amount the data which is stored in hospital information systems (HIS) is hardly directly usable for medical research or clinical benchmarking.

For a proper and scientifically correct analysis of patient from hospital information systems, the relevant data has to be extracted from numerous heterogeneous systems to a common data storage (data warehouse). This extraction usually comes along with a significant reduction of the data, since only selected features of medical cases of interest are extracted. In some cases it might be necessary to manually enrich this electronically stored data with information from handwritten patient records. So, a data storage system is needed, that allows the import of electronical data as well as the user-friendly input of hand-written data. Furthermore, it must be able to semantically and syntactically check the data in order to provide a proper base for further research.

We developed a generic meta-model based data storage system that is able to store data of arbitrary structure. Just by configuration the user is enabled to define the data entities of the system, their attributes and their relations among each other. Furthermore, it is possible to define syntactical and semantical checks to ensure the syntactical correctness and semantical plausibility of the data. So a fully-fledged data storage system including data checks can be setup up for a certain domain within hours.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 40–53, 2012.

In this paper we present the system itself in detail. In a further development step we tried to integrate a clustering algorithm to enable the system to yield similar cases to a current one. Therefore, we evaluated clustering algorithms, according to their usability for similarity search of medical cases. Section 2 describes the system and the meta-model it is based upon. In Section 3 an evaluation of clustering algorithms is presented. In Section 4 we provide first results of our system and Section 5 contains our conclusions and an outlook for further developments.

2 Meta-Model Based Data Acquisition

For a proper and scientifically correct analysis of hospital derived patient data the relevant data, which is usually a small subset of the data stored in a hospital information system, has to be extracted from numerous heterogeneous systems to an common data storage (data warehouse). The data warehouse should not only be able to store the collected data, but should allow the user to create, edit, search, and complete the data records via clearly structured user interfaces, preferably accessible via web for distributed multi-center studies. The research target and medical domain of the study determine the data structure and consequently the implementation of the data storage system. So the whole system (including database, user interface, logical layer) strongly depends on the domain it was implemented for and is hardly reusable in other domains. In order to avoid these dependencies, we developed a web based, highly generic data storage system, that is based upon a generic meta-ER-model.

2.1 Meta ER Model

The Entity Relationship model (ER model) is one of the conceptual models introduced in [2]. However, ER models are also part of this reality and can therefore be described, using an ER model as well. Models of a model are called meta-models. The Object Management Group OMG [3] defines four levels (M_0 - M_3) of meta modeling. Each model at level M_i is an instance of a model at level M_{i+1}. Level M_0 contains the real world or user data. Each object at M_0 is an instance of a model defined in M_1, which is called the model layer. Each model at level M_1 can be seen as an instance of a meta-model at level M_2 - the meta-model layer. Level M_3 contains meta-meta models. A meta-ER model is an ER model describing an ER model. Since ER models are part of the model layer M_1 of the OMG four-level meta-model stack, meta-ER models belong to level M_2. Our M_2 model is able to store the M_1 domain dependent data model as well as the corresponding M_0 data. The core entity of the data model shown in Figure 1 is “Entity”; it basically represents a table in a database. Each *entity* can have *n attributes* (corresponding to the attributes of a table in a database). Foreign key-based relations in a relational database are stored in the “Relation” table. Each *record* represents one single row of a database-table and belongs to one certain *entity.* A “Value” represents the data, which is in each cell of a table.

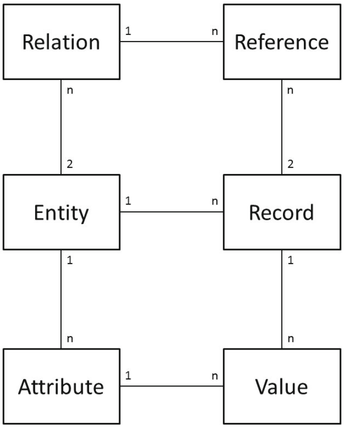


Fig. 1. ER Diagram of our Meta-ER-Data model

Each value belongs to a *record* (row of the table) and an *attribute* (column of the table). Values that represent foreign-key data are stored in the “Reference” table.

2.2 Usage

Before the system can be used to store data, the domain dependent M_1 data model has to be defined by the user. Therefore, the web based administrator back end or a Java based application can be used. The user needs to define the entities of his research domain, their attributes and their relations among each other. Furthermore, the system allows the definition of numerous syntactical and semantic rules in order to ensure the correctness and plausibility of the collected data for subsequent analysis. For all these operations no programming or database skills are needed. Dialogs and Wizard guide the domain- (and not IT-)expert through the setup. The instantiated M_1 model is stored into the M_2 meta-ER-model and is used the create user interfaces like overview tables, input forms, search forms on demand. Changes to the domain specific M_1 data model during the ongoing data acquisition are also possible. The collected data itself is also stored into the meta-ER-model in particular into the tables record, value and reference. When displaying a certain record to the user, the structure of the web page is derived from the meta information stored in the tables entity, attribute and relation, while the content is taken from the corresponding tables record, value and reference.

2.3 Consequences

As stated above, the use of M_2 meta-models instead of M_1 data-models increases the stability of the implemented data structure and source code. Changes in the instantiated data structure can easily be made without manipulating the source code or the database. Another benefit is the high adaptability of the system. M_2 meta-model systems are designed for classes of problems and not for one single problem. They can be reused in many different domains.

Nonetheless, there are some disadvantages as well. Data that can easily be queried out of an M_1 data-model, needs to be assembled by multiple joins out of the meta-model. This is not just an issue when talking about performance, it also complicates the source code. Software developers must get used to think in meta-model terms, which means dealing with another level of abstraction. Especially when it comes to queries meta-models can be very challenging. Another drawback that has to be accepted is the fact that almost all of the data is stored in one single table - the *value* table. This causes an imbalance in the sizes of the tables and leads to one table carrying all the data, while other table remain very small. However, according to our experiences so far, this imbalance did not cause any problems to instances holding up to 20,000 records.

3 Clustering Algorithms

3.1 Motivation

In modern hospital workaday life almost any treatment of a patient is electronically documented. On the one hand a huge amount of medical data is collected, which is mostly used by doctors to get an overview about the current patient's anamnesis or for billing issues. On the other hand, doctors sometimes spend hours to find medical cases that are comparable to their current one. Informations about these similar cases can be used to improve the current patient's treatment and shorten the physician's research time. In order to reliably detect similar cases in the whole dataset a number of clustering-algorithms will be evaluated according to their runtime behavior, "correctness" and their ability to distinguish isolated medical cases from medical cases that are part of a group of similar elements. Furthermore, a proper clustering can be the initial point for further (semi)automated analysis of the stored data (see section 5.2 on page 51).

Clustering algorithms are part of the group of unsupervised learning algorithms. In contrast to supervised learning algorithms, clustering algorithms don't try to forecast a class label for a given vector of numbers, they try to divide the dataset into groups of similar objects. Jain et al define clustering as follows:

> "*Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity.*" [4]

So the intention of clustering is to split up a data set into (meaningful) groups, whereas the samples within a group are more similar to each other than to samples of other groups.

Because of the absence of a class label in the training data there is no way to objectively check the quality of the cluster result. Usually this is done be humans, with respect to the purpose of the data exploration using quality criteria like the Sum-of-Squared-Errors criterion. Hence, the quality of a clustering strongly depends user's point of view and can therefore not be defined objectively. There is no correct or incorrect clustering [5]. For this reason the fully automated use of clustering algorithms is considered problematically. Moreover, many clustering algorithms require the correct number of clusters in the dataset as an input parameter. Often this number is unknown and an object of study as well.

3.2 Preselection of Evaluation Candidates

The clustering algorithm will be integrated in our data acquisition system. Since the domain of application is not known the data itself as well as its structure is unknown as well. Hence, the number of clusters in the dataset is unknown, so a big number of clustering algorithms will not be taken into account, because they take this number as an input parameter. The following clustering algorithms either don't need a number of clusters or calculate this number by their selves.

Consensus Clustering. The clustering method consensus clustering is based on the following assumption:

> "*If the data represent a sample of items drawn from distinct sub-populations, and if we were to observe a different sample drawn from the same sub-populations, the induced cluster composition and number should not be radically different. Therefore, the more the attained clusters are robust to sampling variability, the more we can be confident that these clusters represent real structure.*" [6]

This assumption implies multiple re sampling and clustering of the same data. For each clustering step the stability of the resulting clusters is calculated. The consensus of the best results is presented as a final result.

Self Organizing Maps. Self-organizing maps (SOM) or Kohonen maps are artificial neural networks used for unsupervised learning. The incremental-learning algorithm of SOM arranges the output elements of the neural network in a way that preserves the topological order of the input space. Since the output elements usually are arranged in a two-dimensional regular grid, the SOM performs a mapping from an n-dimensional input space to a two-dimensional output space as well as a clustering of the input datasets. [7]

Nearest Neighbor Approach. The nearest neighbor search yields the n nearest neighbors for a given datapoint. It can also be used to assign a classification label to a yet unclassified sample point according to the class labels of a set of nearest points in the neighborhood [8]. The principal of this approach can be used to not only assign a class label but link the neighboring medical cases to each other. However, a distance boundary is needed to distinguish close - and therefore relevant - medical cases from distant ones.

4 Results

4.1 Test Data

For testing the clustering algorithms four datasets - two generated and two natural - were used. Figure 2 on the next page shows a mapping from their original space into a two-dimensional space using Sammon's mapping algorithm [9] of the four datasets.

(a) iris. In the upper left corner, labeled with character a the iris dataset can be seen. This data set is fairly well-known since it was used by Fisher [10] in several statistical experiments. The data were originally obtained by making four measurements on Iris flowers. These measurements were then used to classify three different species of Iris flowers. Fifty sample vectors were obtained from each of the three species. Thus, the data set consists of 150 points distributed in a 4-dimensional space [9].

(b) gaussian5. The artificially generated dataset b consists of five Gaussian clouds equal in size and shape in a six dimensional space. The clouds are located close to each other, so that overlapping effects occur. Each cloud consists of 50 samples resulting a overall sample count of 250 samples.

(c) aneurysm. The aneurysm dataset consists of 285 samples in a 42 dimensional space. The dataset contains parameters such as size, form, and type of cerebral aneurysm. The visualization of the dataset shows groups of very similar samples as well as isolated and lose connected samples - a very challenging structure for clustering algorithms. The data is provided by the multi-center aneurysm register www.aneurysmen.at which is run by the Landesnervenklinik Linz Wagner Jauregg.

(d) gaussian2diff. This artificially generated dataset consists of 140 samples in a 4 dimensional space, whereas the majority of all samples builds a big cluster and only 20 sample build a smaller one.

4.2 Criteria

Since the structure of the data - and consequently the number of clusters - is completely unknown and the clustering component should work without human supervision, the challenges to this component are rather high. Moreover, projections of the aneurysm dataset (see Figure 2 on the following page) showed a very heterogeneous structure of the test-data. Some samples are located very close to each other and model clusters, while other seem to be more isolated. Based on this situation the following requirements for clustering algorithms were defined:

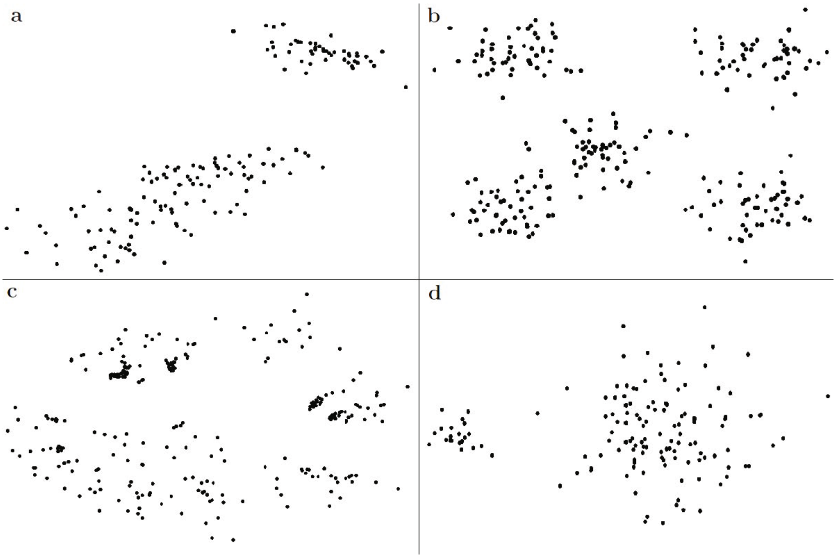


Fig. 2. (a) The iris dataset, (b) Five Gaussian clouds, slightly overlapping, (c) the aneurysm data set, (d) Two Gaussian clouds of different sizes

Discovering of Cluster Numbers. Since the number of clusters in the dataset is unknown the clustering algorithm must be able to determine this number itself or work without it.

Discovering of Isolated Records. Analysis of the aneurysm dataset showed a big number of isolated cases. The clustering component must have the ability to recognize this cases as isolated and not include those cases to the next big cluster.

Result Stability. Many clustering algorithms start from a random initial state. The final result of some of these algorithms (EM algorithm or k-means clustering) strongly depends on the random initialization, while others show a good convergence behavior. Although there have been approaches to improve the clustering initialization for EM and k-means ([11]) the evaluation concentrated on clustering algorithms that either don't start from a random starting point - like hierarchical clustering - or show better convergence behavior.

Updatability. The clustering component will be developed for a web-base storage system for medical data. The data stored in this system is constantly changing and growing. One criteria for evaluating the clustering algorithm is how easy samples can be added or changed, and how reliably they are (re-) inserted into the correct cluster.

4.3 Consensus Clustering

The implementation of the consensus clustering algorithm performs remarkably well in finding the correct number of classes in generated data sets, such as *gaussian5* or *gaussian2diff*. It was still able to determine the correct number of clusters and separate the data clouds from each other even when the distance between the cloud-centers was reduced. Unfortunately, the algorithm failed to split up the aneurysm dataset into a reasonable number of clusters. Figure 3 on the next page shows the clustering result of the aneurysm data set. The algorithm performed well recognizing clearly delimited groups of aneurysms in the lower area of the figure (cluster number 2,7, and 9). In the upper area of the figure where the density of the dataset in lower the separation of the clusters is less clear. It also failed to distinguish very tight grouped aneurysm from their surrounding (cluster 8 in the right area).

Isolated samples are supposed to build a cluster of their own. Consensus clustering tends to attach these cases to an arbitrary cluster. The phenomenon can clearly be seen in the lower right area of Figure 3 on the following page. These isolated cases are labeled with cluster numbers of clusters that are far away from these samples. Analysis of the aneurysms behind these samples showed that these cluster assignments are incorrect.

Due to the fact that consensus clustering is based on multiple clustering of the same dataset the stability of the result is fairly high. Nevertheless, it strongly depends on the number of re-clustering steps, which should not be below 200. Even though the number of clusters and the groups of clustered cases remains constant over several runs of the algorithm, the IDs of the clusters changed. Consensus clustering shows good performance when there are clearly separable clusters, even if their structure is challenging (different sizes of clusters, small inter-cluster distances, etc). The aneurysm data set shows clearly separable clusters in the lower area of the illustration (Figure 3 on the next page) but all samples in the upper regions model a loose cloud of cases. Dividing this area into cluster is hardly possible (even for humans).

The clustering approach in common suffers from a major shortcoming. Figure 4 on page 49 shows the problem. It shows a subset of the clustering result that is shown in Figure 3 on the next page. The sample A and B are located very close to each other; meaning they are very similar. Both are located distant to sample C. Nevertheless, sample B and C are assigned to the same cluster (8), while sample A is member of another cluster (7). So B and C are treated as similar while A and B a treated as they were totally different. This phenomenon appears in the border region of two clusters and it's effect gets amplified when the clusters show a very stretched form. This is not just an issue of consensus clustering it affects all traditional clustering approaches.

Another shortcoming of consensus clustering is the disadvantageous runtime behavior. Consensus clustering is known to be NP-complete. [12] proposes a number of heuristics and refinements to accelerate the clustering but these proposals were not taken into account because the runtime behavior is not the most significant deficit of consensus clustering.

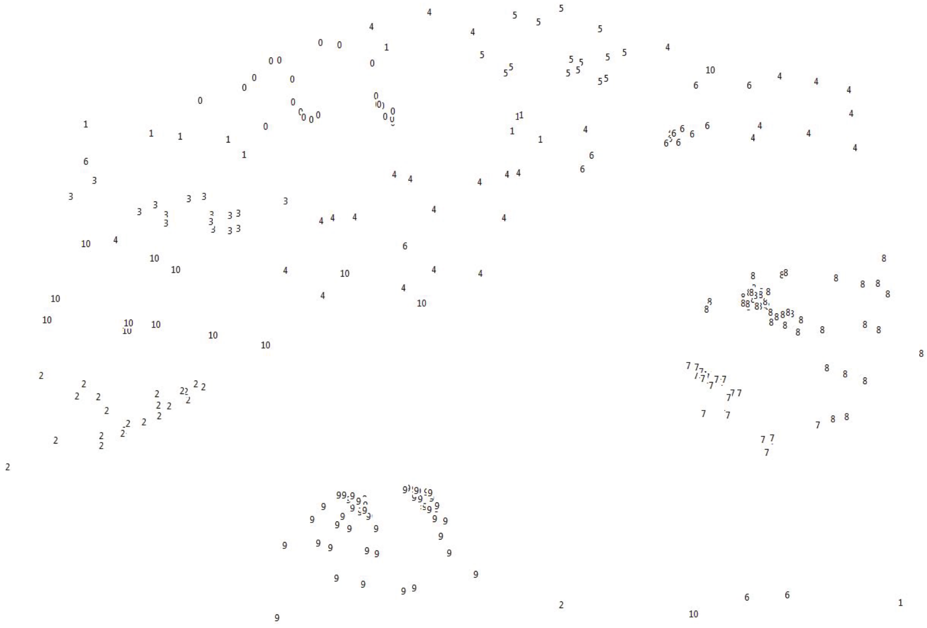


Fig. 3. Result of the consensus clustering algorithm applied to the *aneurysm* dataset

The insertion of a new sample can hardly be done without re clustering the whole data set. One way to avoid a re-clustering is assigning the new the sample to the cluster of his next neighbors; a classic n-nearest neighbor approach [8]. For this procedure the distance between the new sample and all other samples in the data set must be calculated; whereas the calculation time increases with the size of the dataset. In order to avoid this extensive calculations a classifier could be established that assigns the cluster label to new sample, which is also very extensive and error-prone.

4.4 Nearest Neighbor Approach

The nearest neighbor approach is independent from the number of clusters in the dataset. It simply connects those cases, that are close together.

In the initial state of the algorithm all samples are connected. The distance matrix seen as a adjacency matrix results in a fully connected graph. The algorithm thins out the distance matrix and transforms it into a similarity matrix containing only those edges whose length is under a certain threshold. If a sample has no connections to others that are short enough, it loses all of them and remains isolated.

The choice of the distance boundary for similarity is crucial for this algorithm. In case this boundary is defined using a clustering of the distance matrix.

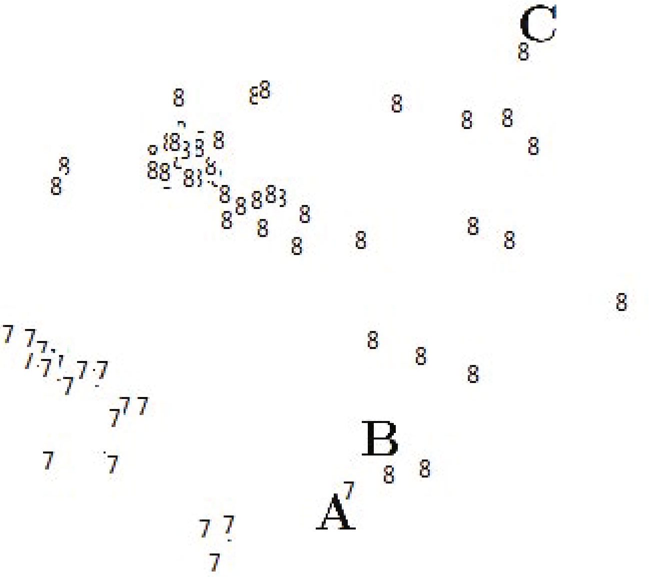


Fig. 4. Problem of clustering

Since the distance matrix is too big to be clustered as a whole, a random subsample is taken and clustered. The randomization and the rather small size of the subsample (for performance reasons) decrease the reliability of this approach.

At a first glance, inserting new samples is a straight forward procedure. Just like the clustering procedure itself, the distances between the new sample and all others are determined. The new record is connected to all old records that are close enough; meaning their inter-sample distance is below the boundary. The catch in this case is that with increasing size of the data set the insertion duration of new sample increases.

A visualization of the nearest neighbor clustering can be seen in Figure 5. While the results for the *gaussian5* data set look very promising the results for the *aneurysm* dataset are not satisfying. Altogether, the visualization of this clustering result is very confusing and strongly depends on the result of the boundary finding.

4.5 Self Organizing Maps

The functional principle of the self-organizing map doesn't require the number of cluster in the data set. It basically performs a mapping from a multi-dimensional feature space to a two-dimensional grid of nodes.

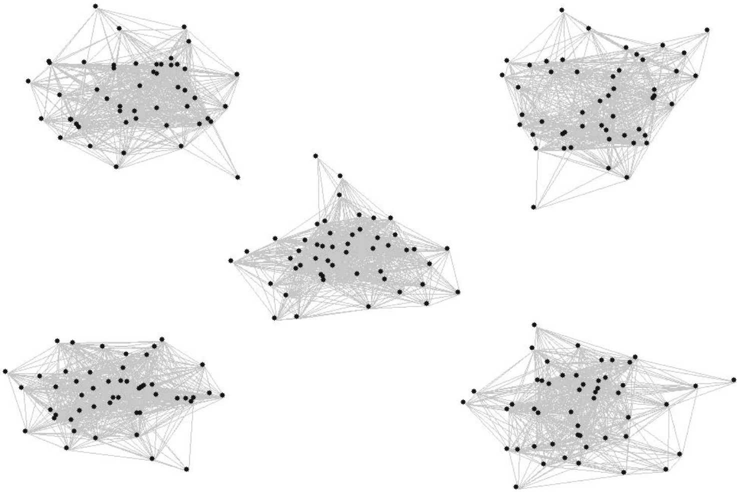


Fig. 5. Graph clustering of the *gaussian5* data set

After the SOM training algorithm has finished each sample is assigned to the node in the grid that has the smallest distance to this sample. Isolated samples are likely to be assigned to a node by their own. So the map is able to recognize isolated cases.

Multiple runs of self-organized map algorithm with the same training data sets show a very high stability of the results. Although the self-organizing map algorithm starts from a random state, the nodes arrange themselves in the same way after each run. Moreover, the groups of samples that were assigned to the same node remained constant throughout the test series.

Inserting a new medical case into the SOM without performing a re-training of the map is straightforward. If the model-vector for each SOM-node is available a new medical case can be assigned to the node that matches it's own feature vector the best. Since the number of nodes in the map is (almost) independent from the size of the dataset, this duration for the insertion doesn't increase with the increasing number of records in the data set. Of course, this insertion doesn't consider the influence, this new case has on the structure of the map. It can be seen as a quick-insert. Given a certain amount of medical cases, the SOM is based upon, the structure will not be influenced significantly by adding one single sample anyway. In a nightly re-training of the SOM the structure will be recalculated, including the values of the new case. Beyond that, the SOM also support updating a medical case without re-training. Just like a new inserted sample is automatically assigned to it's closest node, an edited sample is detached from it's old node and re-inserted into the map.

The self-organizing map offers additional functionality besides it's clustering ability. It is possible to browse through the grid of nodes, to extend the searching area for similar records. Assuming a doctor is looking similar cases for a very extraordinary case. Since, this case is likely to be very isolated it will be assigned to a node of the SOM by it's own. So no similar cases will be found. In this case it is easily possible to extend the search be exploring the neighboring node of the actual SOM-node as well.

5 Summary

5.1 Conclusion

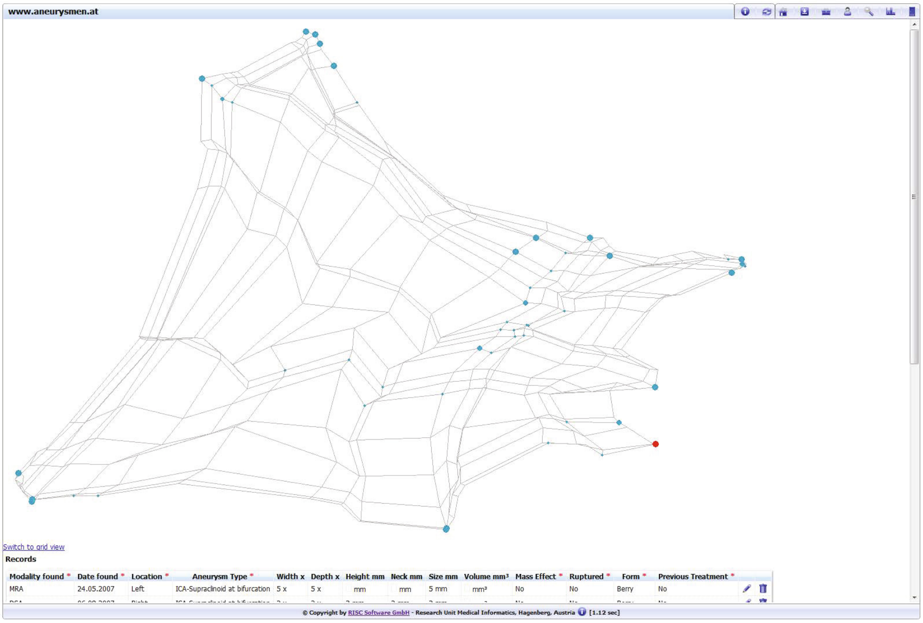
We showed that the most applicable clustering method in this case is the self-organizing map. It fulfills all predefined criteria and outperforms the other clustering algorithms in many ways. Insert and update operations can be done in a constant time; independent from the number of samples in the dataset. The map enables the user to browse from one group of similar samples to next group of samples, whereas both groups are similar to each other - just by exploring the neighboring SOM-nodes. Moreover, visualizing the SOM is more clear-arranged than direct visualization of the data set, especially with increasing size of the data set.

For applying the SOM to a set of data, the user needs to define which attributes of the entities should be taken into account, when calculating the distance measure and how these features are weighted. After that, the basic SOM algorithm is performed and each data set is assigned to a node of the SOM. Figure 6 on the following page shows the visualization of the SOM representing the aneurysm data set. This dynamically created visualization is included into the web interface of the aneurysm register an allows the users to browse the map and look for groups of similar records. After selecting a node of the SOM all patient records assigned to that certain node are listed below the graph. So, the physician can easily find cases that are similar to a current selected.

The system is designed to be set up, maintained and used by non-IT personell. So the barrier to use professional data acquisition systems instead of semi-professional solutions like Excel-sheets or text documents shall be decreased. The integration of algorithms like the SOM takes the same line by allowing users without any IT or algorithmic skills to apply these algorithms to their data.

5.2 Further Research

In the current stage of development the gridsize of the self-organizing map is defined by user configuration. In order to reduce the user input at this place and increase the autonomy of the clustering algorithm dynamic or growing self-organizing map algorithms like, growing self-organizing maps by Alahakoon et al [13], will be evaluated and integrated into the system. Furthermore, both, the basic SOM algorithm as well as growing SOM, still work with numeric input vectors, while the data is stored in XML like tree structures. For the current implementation the tree structured data needs to be denormalized and encoded to a numeric to be precessed by the SOM. In the course of this denormalization structural information gets lost. To counter this drawbacks SOM algorithms for structured data ([14], [15], [16], [17], [18]) will also be evaluated for this project.



www.aneurysmen.at

Switch to grid view

Records

Modality found *	Date found *	Location *	Aneurysm Type *	Width x	Depth x	Height mm	Neck mm	Size mm	Volume mm³	Mass Effect *	Ruptured *	Form *	Previous Treatment *
MRA	24.05.2007	Left	ICA-Supraclinoid at bifurcation	5 x	5 x	mm	mm	5 mm	mm³	No	No	Berry	No

© Copyright by RISC Software GmbH - Research Unit Medical Informatics, Hagenberg, Austria [1.12 sec]

Fig. 6. SOM integrated into the system

The meta information stored in the data model can also be used to support (semi)automated sophisticated statistical analysis like subgroup discovery, regression, classification or pattern recognition. The data acquisition system is going to be extended in a way that it is able to search the data it currently holds for clusters, patterns, interesting subgroups, statistical outliners. Due to the meta information, these kinds of machine-learning and data-mining algorithms can strongly be automatized and offered to users, who are experts in their domain, but in data engineering and machine learning. So the barrier for non-IT personell to apply advanced data-mining algorithms will be relieved. The research, which is described in this paper, builds the base for further ideas which are presented in detail in [19].

Acknowledgment. We would like to thank the Institute for Radiology of the Landesnervenklinik Linz - Wagner Jauregg (Austria) and the Austrian Society for Radiology for providing the aneurysm data set for this study.

References

1. Leiner, F., Gaus, W., Haux, R., Knaup-Gregori, P.: Medical Data Management - A Practical Guide. Springer (2003)
2. Chen, P.P.S.: The entity relationship model - toward a unified view of data. ACM Transactions on Database Systems 1(1), 9–36 (1976)

3. Meta object facility (mof) specification, OMG-Document ad/97-08-14 (September 1997)
4. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31 (1999)
5. Elhawary, M., Nguyen, N., Smith, C., Caruana, R.: Meta clustering. In: Sixth IEEE International Conference on Data Mining, vol. 1, pp. 107–118 (2006)
6. Monti, S., Tamayl, P., Mesirov, J., Golub, T.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52, 91–118 (2003)
7. Kohonen, T.: The self-organizing map. Neurocomputing 21, 1–6 (1998)
8. Cover, T.M.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13, 21–27 (1967)
9. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Transactions on Computers 18, 401–409 (1969)
10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, 178–188 (1936)
11. Bradley, P.S., Fayyad, U.M.: Refining initial points for k-means clustering. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 91–99 (1998)
12. Goder, A., Filkov, V.: Consensus clustering algorithms: Comparison and refinement. In: Proceedings of the Workshop on Algorithm Engineering and Experiments, pp. 109–117 (2008)
13. Alahakoon, D., Halgamuge, S., Srinivasan, B.: Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Transactions on Neural Networks 11(3), 601–614 (2000)
14. Hagenbuchner, M., Sperduti, A., Tsoi, A.C.: A self-organizing map for adaptive processing of structured data. IEEE Transactions on Neural Networks 14(3), 491–505 (2003)
15. Hammer, B., Micheli, A., Sperduti, A., Strickert, M.: A general framework for unsupervised processing of structured data (2004)
16. Hagenbuchner, M., Tsoi, A.C.: A supervised training algorithm for self-organizing maps for structures. Pattern Recognition Letters 26(12), 1874–1884 (2005)
17. Hagenbuchner, M., Sperduti, A., Tsoi, A.C., Trentini, F., Scarselli, F., Gori, M.: Clustering XML Documents Using Self-organizing Maps for Structures. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 481–496. Springer, Heidelberg (2006)
18. Martín-Merino, M., Muñoz, A.: Extending the SOM Algorithm to Non-Euclidean Distances via the Kernel Trick. In: Pal, N.R., Kasabov, N., Mudi, R.K., Pal, S., Parui, S.K. (eds.) ICONIP 2004. LNCS, vol. 3316, pp. 150–157. Springer, Heidelberg (2004)
19. Girardi, D., Dirnberger, J., Giretzlehner, M.: Meta-model based knowledge discovery. In: 2011 International Conference on Data and Knowledge Engineering (ICDKE), pp. 8–12 (2011)

A Mobile Based Authorization Mechanism for Patient Managed Role Based Access Control

Cátia Santos-Pereira[1], Alexandre B. Augusto[2], Manuel E. Correia[2,3],
Ana Ferreira[1,4], and Ricardo Cruz-Correia[1]

[1] Center for Research in Health Technologies and Information Systems (CINTESIS),
Faculty of Medicine of University of Porto (FMUP), Portugal
[2] Center for Research in Advanced Computing Systems (CRACS),
Department of Computer Science, Faculty of Science of University of Porto, Portugal
[3] Department of Health Information and Decision Sciences (CIDES), FMUP,
Portugal
[4] Informatics Centre, FMUP, Portugal
{catiap,amlaf,rcorreia}@med.up.pt, {aaugusto,mcc}@dcc.fc.up.pt

Abstract. The Internet has proved the enormous benefits that can be accrued to all players involved in online services. However, it has also clearly demonstrated the risks involved in exposing personal data to the outside world and constitutes at the same time a teeming breeding ground of innovation for highly flexible security solutions that can minimize these risks. It is now widely believed that the benefits of online services to healthcare in general supplant the risks involved, provided adequate security measures are taken and the role played by all the parties involved, be they physicians, nurses or patients are clearly outlined. Due to the highly sensitive nature of the data held on the Electronic Health Record (EHR), it is commonly agreed that providing online access to patients EHR to the outside world carries an unacceptable level of risk not only to the patients but also to the healthcare institution that plays a custodian to that sensitive data. However, by sharing these risks with the patients, healthcare institutions can start to equate the possibility of providing controlled exterior online access to patients EHR. The mobile phone is nowadays the preferred mean by which people can interact with each other at a distance. Not only that, the smartphone constitutes the full embodiment of the truly personal device users carry constantly with them, everywhere. They are therefore the ideal means by which the user can casually and conveniently interact with information systems. In this paper we propose a discretionary online access rights management mechanism based on the Role Based Access Control (RBAC) model that takes advantage on the personal/technical characteristics and data communications capabilities of the smartphone in order to provide patients with the means by which they can conveniently exercise safe discretionary online access permissions to their own EHR.

Keywords: Patient Empowerment, e-health, Electronic Health Records, RBAC, Secure Mobile Wallet, PKI, smartcard, QR codes and Secure Tokens.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 54–68, 2012.

1 Introduction

Nowadays, patients want to be better informed about their medical conditions and play a more active role on their own treatments. They usually consult online information by using search engines to educate themselves about etiology, treatments, and the prognosis of the medical conditions. Getting access to their medical records would help the patient to better understand their medical conditions [1]. On this issue the European Recommendation [2] and American Legislation [3] for protection of medical data agree that the patient must have access to his/her medical record and play a major role in the decisions regarding the content and the distribution of his/her medical data [4].

Currently, for patients to have access to their medical records they need to write a request to their custodian healthcare institution and the response delay depends on their country legislation (e.g. 10 days in Portugal [5], 21 days in England [6] and 30 days in USA [3]). We believe that the latest developments in digital communications and information technologies should provide the patients a simpler and more secure way to access their medical records and at the same time provide a better collaboration and interaction experiences with the healthcare professionals [7].

In the healthcare domain, patients digital data is normally collected into what is called the Electronic Health Record (EHR). The EHR encompasses many functions that can include different types of data items such as diagnoses, medications and operations [8,9]. The EHR is nowadays indispensable for health institutional purposes and could be used to empower patients by giving them the necessary information to play a more active role in their own health and in their families health as well [10].

Unfortunately, by opening up the access to the EHR with inadequate access control mechanisms and policies carries some substantial risks as illustrated by the grim statistics observed during the period of 2006 to 2007, where in the USA, over 1.5 million names were exposed during data breaches that occurred in hospitals [11]. One of the most important and complex requirement for eHealth systems [12] is to keep patient's information private and secure. The EHRs are daily accessed by a diversity of health professionals that have different objectives according to their functions. Appropriate access control mechanisms and policies are essential to provide a good balance between usability and confidentiality. These procedures constitute the core of the authorization process on eHealth systems, in other words, they are responsible to manage the EHR access by granting access only to previously authorized persons [13].

The patient authorization model proposed by *Santos-Pereira et al.* [7] is to be used and customized by the patient. This model effectively combines the characteristics of Role Based Access Control (RBAC) model [14], ISO 13606-4 [15], temporal constraints (GTRBAC) [16] and break-the-glass mechanism (BTG-RBAC) [17]. The access permissions of a role to a specific EHR component is dependent on the previously mapping made by the administrator of the model (usually the patient). A customized role can have access to an EHR record component if the administrator defines any of the create, read, update, delete or break-the-glass operations to be part of the record access permissions.

Tacconi et al. [18] states that smartphones are revolutionizing many sectors and aspects of our economy, including social networks and healthcare. Based on this we decided to employ the smartphone as the tool to establish a more interactive relationship between the system and its users, since the smartphone by its own very nature as a personal communication device usually follows their owners everywhere. They constitute an ideal platform to develop and provide any-time and any-where fast user interactions.

Our aim is to define a patient-centric infrastructure to manage medical data access in a reliable and secure way. To realize our objective we rely on the: (a) patient authorization model defined by *Santos-Pereira et al.* [7] to define patient centered access control and administration; (b) Extensible Messaging and Presence Protocol (XMPP) to establish the communication between the mobile devices and the healthcare institutions; (c) usage of dynamic web services in order to create the necessary communication nodes. To this infrastructure we call OFELIA (Open Federated Environments Leveraging Identity and Authorization). OFELIA provides the means to build a secure web based service infrastructure where the patient can access and customize access permissions to his/her own EHR in a complete patient-centric way using his own smartphone as an authorization broker [19].

In OFELIA access control is exercised by the means of a secure mobile identity digital wallet, secured by a Public Key Infrastructure (PKI) with keying operations provided by a smartcard for mobile devices. This secure mobile identity wallet allows the patient to exercise a flexible based access control over their EHR thus disclosing its data only to pre-authenticated and previously authorized users, for certain well defined periods of time at the data owners discretion.

The rest of the paper is organized as follows. In Section 2, we review the system mechanisms and technologies, describing how they constitute the OFELIA architecture. In Section 3 we describe the necessary steps to establish a trust connection in order to request an EHR access. In Sections 4 and 5 we present an usage case scenario and then discussed some issues and their possible solutions. In Section 6 we present a preliminary conclusion about our proposed architecture and delineate our plans for future work.

2 Security Mechanisms and Technologies

In this section we present the mechanisms and technologies employed to implement the OFELIA authorization infrastructure. Each mechanism/technology is presented in some detail and then we explain how the functionalities can be integrated to the proposed authorization model.

2.1 Patient Authorization Model

Santos-Pereira et al. [7] proposed an authorization model where the concept of Patient Healthcare Network (PHN) is defined and is composed by all the healthcare institutions that the patient may attend and where his medical records are kept (e.g. hospitals and healthcare facilities). The knowledge of which healthcare institutions belong to the patient's PHN is very important, because if the patient wishes to access his medical records he should have been previously enrolled within all these institutions. In our vision each healthcare institution deploys its institutional EHRs together with an OFELIA web service that is described in subsection 2.7. The concept of PHN and the need for the patient to access his medical information, within multiple healthcare institutions, is very important however is not contemplated in this paper because we are only focusing in the authorization process architecture. Nevertheless, the PHN concept and implementation in all its extent is a fundamental step to be addressed as future work.

Santos-Pereira et al. model integrates a set of functional roles that categorizes the accesses to the patient EHR into three main groups: subject of care (SC) (Group I), healthcare professionals (Group II) and administrative staff (Group III) (see Figure 1).

In this work we focused in Group I to explore the idea of patient empowerment by deploying an infrastructure that allows patients to access and share their EHR record components based on functional roles.

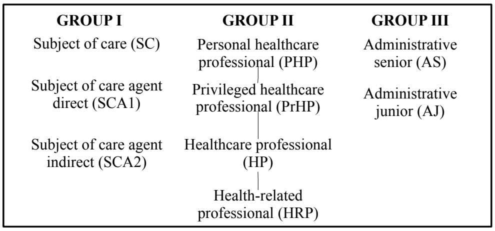

GROUP I
Subject of care (SC)
Subject of care agent direct (SCA1)
Subject of care agent indirect (SCA2)
GROUP II
Personal healthcare professional (PHP)
Privileged healthcare professional (PrHP)
Healthcare professional (HP)
Health-related professional (HRP)
GROUP III
Administrative senior (AS)
Administrative junior (AJ)

Fig. 1. Functional Roles groups and Hierarchies [7]

2.2 Quick Response Code

The Quick Response codes (QR codes) are two-dimensional square shapes that encode a reasonable amount of digital information (several Kilobytes of chars) in a small amount of space. The encoding is achieved with the careful positioning of varying size black and white smaller squares within the 2D space defined by the QR square. These 2D codes are normally displayed within web pages or printed in paper posters and are employed to quickly exchange digital information with mobile devices that would otherwise had to be entered by hand. This is accomplished by having the mobile device to digitally scan and decode the displayed QR code with its built-in optical camera [20].

In OFELIA, QR codes are displayed at computers displays for an auto-enrollment process of smartphones into the healthcare institutions. QR codes are a very convenient way of conveying a reasonably amount of secret shared information to a smartphone that would otherwise be very cumbersome to input by hand by the user. The usage of QR codes to share secret information between ehealth systems and smartphones, can in a way, be seen as the establishment of a rather new special security layer by taking advantage of the analog security properties of the optical channel that is employed during the scanning of the QR codes by the smartphone. In other words, the QR codes can be used to simplify and make practical the enrollment process between the OFELIA web service (within the healthcare institution) and the user smartphone.

2.3 Extensible Messaging and Presence Protocol

The Extensible Messaging and Presence Protocol (XMPP) is a widespread open technology, employed for almost real-time messaging style communication, that takes advantage of the eXtensible Markup Language (XML) as a base format for exchanging information [21]. XMPP provides a complete standard set of services [22] like certificate based authentication and asynchronous one-to-one and many-to-many messaging services that are extensively employed in our propose as the network transport layer infrastructure.

Arguably, in the mobile world an implicit direct Internet communication with a personal device is generally not possible due to the shortage of public IP (*Internet Protocol*) addresses faced by Internet service providers. In the near future, the IP version 6 (IPv6) is supposed to solve this problem, however we believe that the mobile telecommunications operators will not allow for directly addressable mobile devices from the Internet due to their less flexible business plans which regard mobile devices, smartphones in particular, as a strict consumer device, not as a service provider.

Towards this end, XMPP is proving to be an almost ideal communication infrastructure for our propose to circumvent these communication restrictions because of its ability to efficiently operate over HTTP (*Hypertext Transfer Protocol*) by the means of the BOSH (*Bidirectional-streams Over Synchronous HTTP*) [23] protocol, where two non directly addressable devices, located on private closed intranets and with minimal Internet access, can locate each other over the Internet and then freely exchange messages between themselves in a reliable and secure way [24].

2.4 The Trust Infrastructure

The management of trust between the healthcare institutions and the smartphones is essential to our scenario. To establish this trust we rely on a Public Key Infrastructure (PKI) that is responsible for the management of the certificates that are at the core of the privacy, trust, non-repudiation and authentication infrastructure mechanisms that we need to put in place to secure our architecture.

To establish a stronger and therefore more trustworthy identity and authentication between the different actors (personal smartphones and healthcare institutions), we rely on the deployment of a well managed standard compliant PKI that can also sign PGP (*Pretty Good Privacy*) and X509 certificates. These certificates are then used as securely vouched identity credentials that can be employed to establish highly secure communication channels, with a reasonable degree of non-repudiation properties and trust between the parties involved in the communication.

2.5 MicroSD Mobile Security Card

Due to the pivot role played by the smartphone in our vision, it is vital to guarantee a more trustful patient identity and assure strong authentication for the communication mechanisms. A more traditional file based keystore to protect the keys of the identity certificates would not be secure enough because a regular file can be easily copied and the smartphone can be a target of attacks where this keystore file can be compromised. It is reasonable to put the encrypted file based keystore security in tandem with the security provided by a much simpler login/password based scheme. In fact, an attack on a password protected keystore file involves a password guessing attack completely analogous in terms of complexity to what happens with an attack directed towards a login/password scheme.

To solve this weakness we rely on the security properties of the mobile security card (MSC) [25] for mobile devices. These security cards are composed by a flash memory and a smartcard component that provides the necessary crypto components and device physical non tampering security features [26]. This allows us to guarantee a two factor authentication required for sensitive data exchange scenarios. To also encourage and provide for a greater level of user responsibility and trust in the system, we also employ identity certificates (X509, PGP) [27] for MSC internally generated crypto-key pairs. These certificates are signed by the users citizen card (eID) [28] to further ascertain the smartphone authenticity to the network and by the healthcare institution to establish the possibility of federation between healthcare institutions in a future work.

2.6 The OFELIA Secure Access Authorization Token

An authorization token (AT) can be seen as a secure digital object that an authorized person needs to present in order to have direct access to another person's resources. In other words the authorization token looks like a valet key for data access, the one who possesses the key has temporary restricted access to the valet key emitter data.

These authorization tokens are also very hard to falsify and take the form of a small base64 encoded XML excerpt, containing elements for a large pseudo-random number [29], and a simple statement describing the authorization validity restrictions which apply to a particular authorization. This statement can express for example temporal restrictions. The XML excerpt is then digitally signed by the smartphone MSC private key and the resulting XML document is then encoded into a base64 string which constitutes the OFELIA authorization token.

The ATs provide a flexible security mechanism for EHR access control, allowing the EHR requesters a restricted and controlled access to the selected EHR record components and discarding the necessity of sharing and managing other types of credentials like login/passwords. In our scenario, these ATs are directly linked with a specific access role, defined by the EHR access control administrator, the subject of care, to the requester at the moment of the authorization. These tokens are only shared with the requesting healthcare institution since the EHR requesters are always associated with a specific identity. It is also important to clarify that in our vision the EHR access control administrator, the subject of care, also maintains the revocation rights by being able to unconditionally revoke these tokens at any given moment.

2.7 OFELIA Web Service

As illustrated in Figure 2, the OFELIA web service (OWS) is responsible for the management of the EHR authorization access process. The OWS is structured into three main component nodes: a XMPP server, an External Web Application (EWA) and an Internal Management Service (IMS).

The XMPP server grants a strong patient authentication method due to the mandatory usage of PGP certificates at the login process and a secure and asynchronous communication between smartphones and the EWA. The patient enrollment process with the XMPP server is illustrated by Figure 3 on step 4.

The External Web Application (EWA) is the service responsible for answering the HTTP(s) external requests, in other words, every time a patient wants to access his or other previous authorized EHR data, he must make a request to the EWA by using a computer with Internet access. When requested the EWA replies with an invalid session key encoded as a QR-code, which is then scanned and decoded by the patient smartphone that in turn sends a request to the XMPP server to validate the session key. This process is presented in Figure 5 in steps 1 to 4. After the patient's request, the XMPP server communicates with the EWA confirming the patient authenticity and requests the session key validation by presenting the patient certificate. Now the EWA can validate the session key by annexing the appropriate identity to it and finally establishing a valid session to the patient.

The Internal Management Service (IMS) is responsible to: query the healthcare institution's EHR based on the requester role that was previous established on the RBAC model by the Subject of Care at the moment of the authorization process; manage the patient enrollment into the healthcare institution by directly communicating to the registration healthcare institution computer; and to manage the patient identity and his authorization tokens. So after a patient successfully authenticates based on his identity permissions the IMS sends the information of what EHRs are accessible to the patient and respective pre-defined roles are linked to each one of these EHRs. This procedure is illustrated in Figure 5 step 5. This process only involves the IMS and the healthcare institution EHR database since the ATs are stored in the IMS.

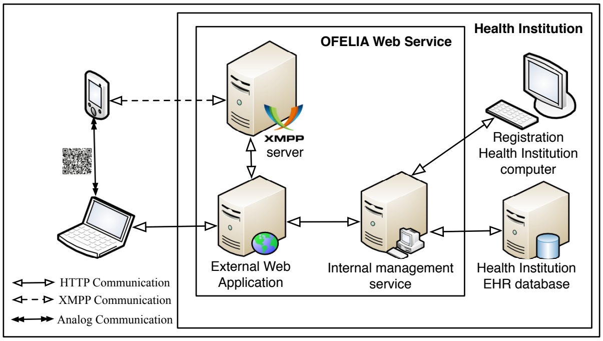


Fig. 2. Patient centric authorization architecture

2.8 OFELIA Mobile Application

The OFELIA Mobile Application (OMA) can be seen as an identity digital wallet to be deployed by the smartphone that is responsible to: strongly prove the patient identity within the healthcare institutions; create and manage the authorization tokens by providing the necessary means to its authorization and revocation; be the "unbreakable" bridge between the subject of care and the healthcare institution by constantly following their owners everywhere as the "de facto" personal mobile device.

To fulfill these objectives, OMA is composed by a database to store the ATs information, a MSC to guarantee patient's identity and authentication, and three service libraries: a XMPP client connector to establish the communication, the OFELIA secure access authorization token to provide the methods to generate the ATs in a secure way, and a QR-code library to read and interpret the QR-codes.

3 Enrollment and Authorization Processes

In this section we describe the enrollment process of a patient to a healthcare institution and all the necessary steps to realize the authorization process in order to gain access to his/others EHR record components.

3.1 Patient's Healthcare Institution Enrollment

Due to the patient's low knowledge about privacy, security and IT issues, this enrollment is done within his health institution provider, resulting in a physical security layer done by the responsible support member in service. Figure 3 presents the 6 steps to establish patient enrollment in a healthcare institution:

Step 1: The subject of care (the patient) accesses the registration healthcare institution computer and authenticates himself by using his citizen card (eID) on the pin pad machine of the registration healthcare institution computer.

Step 2: A pair of QR codes is presented one by one by the registration healthcare institution computer. The patient now reads the first QR code with his smartphone for the installation of the OFELIA Mobile Application (OMA) (could be skipped if the user have already installed OMA). Then, by using the OMA, the patient reads the second QR code for an auto-enrollment on OMA of his healthcare institution. This second QR code brings 3 components: an OFELIA web service (OWS) XMPP credential composed by a JabberID and a password, the OWS XMPP address configuration and a session key for the establishment of a session with the OWS.

Step 3: The OMA accesses the OWS XMPP by authenticating with its PGP certificate (generated by the OMA using the MSC). It then sends the previously exchanged session key obtained from the QR code in order to establish a link between the mobile session and the registration healthcare institution computer. Now the OWS pre-registers the patients certificate.

Step 4: To verify the patient authenticity the OWS sends via XMPP to the OMA a four digit one-time password encrypted with the patient pre-registered PGP certificate.

Step 5: The OMA decrypts the one-time password and presents it on the smartphone monitor requesting the patient to handily insert it on the registration healthcare institution computer. This process guarantees that it was the patient's smartphone who read the QR-code, in other words, the correct patient PGP certificate was exchanged.

Step 6: The OWS finishes the registration by sending via XMPP to the OMA the patient PGP certificate signed by the patient's citizen card (eID) and the healthcare institution itself. This PGP certificate is stored by OMA.

It is important to understand that this process of double signature grants a high level of identity and authenticity. The citizen card (eID) is issued by the government of the patient's country and its signature grants our proposal a real civil identity. The healthcare institution's signature is used to prove the patient's enrollment entity.

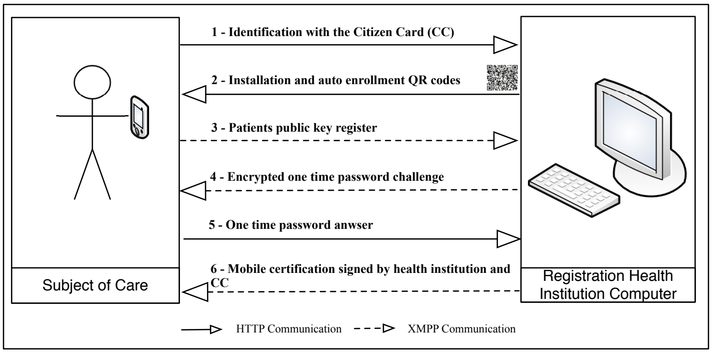

1 - Identification with the Citizen Card (CC)
2 - Installation and auto enrollment QR codes
3 - Patients public key register
4 - Encrypted one time password challenge
5 - One time password anwser
6 - Mobile certification signed by health institution and CC
Subject of Care
Registration Health Institution Computer
HTTP Communication
XMPP Communication

Fig. 3. Subject of care enrollment to a healthcare institution

3.2 EHR Authorization Request

In order to obtain access to other patient's EHR an authorization request must be triggered by a requester (e.g. subject of care agent direct or indirect) who was previously enrolled to the healthcare institution holding that EHR. This whole process is done by using XMPP communication and is described in Figure 4 in 4 steps:

Step 1: The requester, by using the OMA on his smartphone requests the OWS an EHR authorization access by inserting the desired patient JabberID.

Step 2: The OWS sends the EHR authorization request to the OMA of the EHR access control administrator, the subject of care (SC), including the descriptive information about the requester.

Step 3: The EHR access control administrator (SC) is notified on his smartphone by the OMA and based on the requester descriptive information he has to decide. If the subject of care agrees to give any access to his EHR record components, he has to select which functional role the requester will be attributed or to define a more specific role by subscribing any other functional role[7]. After the owner's authorization, the OMA generates an authorization token (AT), signs it and sends it to the OWS that stores the AT for possible later revocation.

Step 4: The OWS sends a report answer with the detailed information (the attributed role and the authorization expire date) to the OMA requester.

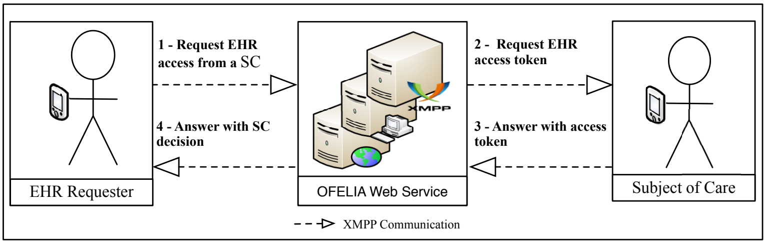

1 - Request EHR access from a SC
4 - Answer with SC decision
2 - Request EHR access token
3 - Answer with access token
XMPP
EHR Requester
OFELIA Web Service
Subject of Care
XMPP Communication

Fig. 4. Subject of care EHR authorization access request

3.3 EHR Access Request

The process to request an already authorized EHR data can be done by any computer with Internet access plus the usage of the user's smartphone to handle the authentication process. As we can see on Figure 5 this process is described in 7 steps:

Step 1: The user requests, via a web-browser, the OFELIA healthcare institution website.

Step 2: A QR-code with an invalid session key is returned, in other word a session key without access permissions is returned.

Step 3: The user, using the OMA on his smartphone, reads the QR-code that contains the session key.

Step 4: The OMA sends via XMPP to the OWS the invalid session key signed with the PGP certificate from the mobile secure card (MSC). It is important to understand that the XMPP authentication method provides a strong non-repudiation method since the requester's PGP certificate is validated during the login process.

Step 5: The OWS links the presented user's identification to the session key and based on the requester access authorization tokens, returns a list of patients and their roles to the requester's computer web browser. Now the session key is validated on OWS.

Step 6: The requester by using his web browser, consults the returned list and chooses the patient he wishes to consult by sending a request to the OWS.

Step 7: The OWS returns to the requester's computer web browser the selected patient's EHR data based on the requester's authorized role.

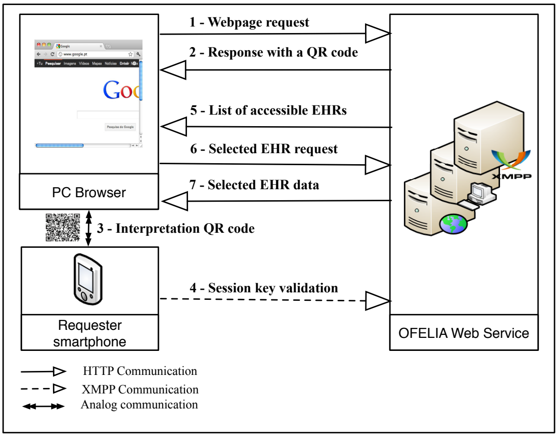


Fig. 5. Subject of care EHR access request

4 Storyboard

To better understand the capabilities provided by our authentication and authorization architecture, we have defined a storyboard to exemplify how the presented patient centric authorization architecture behave on a real healthcare scenario.

Katherine, a 50 years old woman that resides in Mystic Falls, has recently finished her radiotherapy treatments after being diagnosed with breast cancer. Her daughter, Agnes, who lives in a different city 400km away, desires to monitor her mother's follow-up consultations.

Assuming that both mother and daughter are already enrolled at the same healthcare institution, Agnes, the EHR requester, requires Katherine, the subject of care, an authorization to access Katherine's EHR with her smartphone as the subject of care agent direct functional role. Katherine, also using her smartphone decides to grant access to her daughter. However, Katherine wants to customize some of the access control rules of the subject of care agent direct functional role since she desires to omit the treatments' record component, creating the specific role "Patient's Daughter" for that purpose. Now Agnes accesses her healthcare institution website (that triggers the External Web Application from the OWS) with her browser and a QR code is returned and read by the OMA that handles the authorization process with the XMPP server into the OWS. After that, Agnes browser automatically refreshes with a list of the patients for which she has permissions to access. Now Agnes selects her mother assuming her assigned role, "Patient's Daughter", allowing Agnes to read the wished follow-up consultations record component.

Figure 6 illustrates the use-case of the above described storyboard. This use-case shows an example of the EHR folder regarding the Breast Pathology components [15]. The user Agnes accesses Katherine's EHR, with the "Patient's Daughter" role, attributed by her mother, which gives Agnes permissions to only read the following components: *demographic data*, *family history*, *consultations* and *complementary diagnostics tests*. Due to the role restrictions made by her mother, Agnes cannot access the *treatments* component.

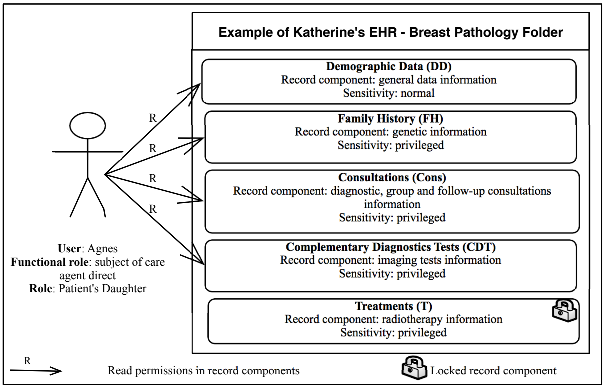

Example of Katherine's EHR - Breast Pathology Folder
Demographic Data (DD)
Record component: general data information
Sensitivity: normal
Family History (FH)
Record component: genetic information
Sensitivity: privileged
Consultations (Cons)
Record component: diagnostic, group and follow-up consultations information
Sensitivity: privileged
Complementary Diagnostics Tests (CDT)
Record component: imaging tests information
Sensitivity: privileged
Treatments (T)
Record component: radiotherapy information
Sensitivity: privileged
R
R
R
R
User: Agnes
Functional role: subject of care agent direct
Role: Patient's Daughter
R
Read permissions in record components
Locked record component

Fig. 6. Use Case related to the storyboard

5 Discussion

Regarding the proposed architecture, the storyboard and use case present a common scenario where a patient's relative, in this case a daughter who is far located from her mother, wants to follow her mother's consultations. This scenario shows how simple it can be for the patient's daughter to request and obtain access to her mother's EHR. Due to the flexibility of our architecture, any previous enrolled user can easily request for an EHR access depending only on the acceptance of the access control administrator, usually the patient.

Despite the annoyance caused by the initial healthcare institution enrollment and the difficulties associated with the usage of our model mechanisms, we believe our approach has dealt with the persistent problem of the patients EHR access. Not only for solving the long wait for the EHR access but also for solving the problem of the outdated EHR record components since the access in our model is processed as requested. In other words the patient can access and administer at any moment and anywhere the most recent EHR record components.

However, in some cases the patients do not have the capability to administer their own EHR access control, due to problems like minimal required age or mental illness. To circumvent this kind of problems our model suggests an enrollment of a legal guardian with the functional role of *subject of care agent direct* granting the legal guardian full access control over the ward's EHR.

Since the smartphone is the key for opening the access to the EHR, its loss or malfunction could bring some substantial problems. These types of problems are usually related with the user's identity proof in order to allow the re-enrollment on the healthcare institution and the certificate revocation. To solve this identity issue, we relied on the signature functionality of the citizen card since this is emitted by the patient governmental entity. This signature usage assures a strong identity proof in order to allow a revocation or a re-enrollment process.

The trust between healthcare institutions is a real problem, bringing communication limitations into healthcare information systems. This weakness limits the users to their own healthcare institution, in other words, it is not possible for two different users enrolled within different healthcare institutions to give authorization permissions between them.

6 Conclusion

With the enormous growth of technologies, the world legislation concern about health data access and the arising of patients' interest to be in control of their medical records, the authors feel it is the right time for an architecture that can give the patients the means to securely and easily access and define access control permissions to their medical records. Our proposed architecture provides for this need by granting the necessary security means as well as promoting the patient empowerment concept.

In order to allow the patient to access and manage his medical data, future work includes the implementation and evaluation of our proposed architecture within a specific case study in a real healthcare institution, more precisely on São João hospital centre, which is the second biggest hospital in Portugal; and a research about federation networks in order to solve the problem of the communication trust between healthcare institutions, as already mentioned in section 5.

Acknowledgments. This work is funded by the ERDF through the Programme COMPETE and by the Portuguese Government through FCT - Foundation for

Science and Technology, project Open Federated Environments Leveraging Identity and Authorization [PTDC/EIA- EIA/104328/2008]. and is being conducted with the institutional support provided by FMUP and DCC/FCUP and the facilities and research environment gracefully provided by the CINTESIS and CRACS (Center for Research in Advanced Computing Systems) research unit, an INESC TEC associate of the Faculty of Science, University of Porto.

References

1. Ebadollahi, S., Coden, A.R., Tanenblatt, M.A., Chang, S.-F., Syeda-Mahmood, T., Amir, A.: Concept-based electronic health records: opportunities and challenges. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, MULTIMEDIA 2006, pp. 997–1006. ACM, New York (2006)
2. Council of Europe. Protection of medical data - recommendation no r (97) 5 (1997)
3. U.S. Department of Health & Human Services. Health insurance portability and accountability act (1996)
4. Pereira, C., Oliveira, C., Vilaa, C., Ferreira, A.: Protection of clinical data - comparison of european with american legislation and respective technological applicability. In: HEALTHINF 2011, pp. 567–570 (2011)
5. Republica Portuguesa. Lei acesso aos documentos da administraçao 46/2007 (2007)
6. NHS choices. How do i access my medical records (health records)?, 15/09/2010 (2012)
7. Santos-Pereira, C., Antunes, L., Cruz-Correia, R., Ferreira, A.: One way to patient empowerment - a proposal for an authorization model. In: Proceedings of the HealthInf 2012 - International Conference on Health Informatics, pp. 249–255 (2012)
8. Hyrinen, K., Saranto, K., Nyknen, P.: Definition, structure, content, use and impacts of electronic health records: A review of the research literature. International Journal of Medical Informatics 77(5), 291–304 (2008)
9. Peleg, M., Beimel, D., Dori, D., Denekamp, Y.: Situation-based access control: Privacy management via modeling of patient data access scenarios. J. of Biomedical Informatics 41(6), 1028–1040 (2008)
10. Dept. of Health & HS. The office of the national coordinator for health information technology (2011)
11. Kroll Fraud Solutions. Healthcare information and management systems society (himss) analytics report: Security of patient data. Technical report, Kroll Fraud Solutions (2008)
12. Watts, J., Yu, H., Yuan, X.: Case study: Using smart cards with pki to implement data access control for health information systems. In: IEEE Southeastcon 2010: Energizing Our Future, pp. 163–167 (2010)
13. ISO/TS 22600-2. Health informatics - privilege management and access control (2006)
14. Kuhn, R., Ferraiolo, D., Sandhu, R.: The nist model for role-based access control: towards a unified standard. In: Proceedings of the Fifth ACM Workshop on Role-Based Access Control, pp. 47–63 (2000)
15. CEN/ISO EN 13606-4. Health informatics - electronic health record communication - security (2009)

16. Joshi, J.B.D., Bertino, E., Ghafoor, A.: Temporal hierarchies and inheritance semantics for gtrbac. In: Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies, SACMAT 2002, pp. 74–83. ACM, New York (2002)
17. Ferreira, A., Chadwick, D., Farinha, P., Correia, R., Zao, G., Chilro, R., Antunes, L.: How to securely break into rbac: The btg-rbac model. In: Proceedings of the 2009 Annual Computer Security Applications Conference, ACSAC 2009, pp. 23–31. IEEE Computer Society, Washington, DC (2009)
18. Tacconi, C., Mellone, S., Chiari, L.: Smartphone-based applications for investigating falls and mobility. In: Proceedings of the International Conference on PervasiveHealth and Workshops 2011, pp. 258–261 (2011)
19. Augusto, A.B., Correia, M.E.: OFELIA – A Secure Mobile Attribute Aggregation Infrastructure for User-Centric Identity Management. In: Gritzalis, D., Furnell, S., Theoharidou, M. (eds.) SEC 2012. IFIP AICT, vol. 376, pp. 61–74. Springer, Heidelberg (2012)
20. Huang, H.-C., Chang, F.-C., Fang, W.-C.: Reversible data hiding with histogram-based difference expansion for qr code applications. IEEE Transactions on Consumer Electronics 57(2), 779–787 (2011)
21. Saint-Andre, P., Kevin Smith, A., Remko Tronon, A.: XMPP: The Definitive Guide Building Real-Time Applications with Jabber Technologies. O'Reilly Media, Inc. (2009)
22. Saint-Andre, P.: Xmpp: Core. RFC 3920, IETF (2004)
23. Paterson, I.: Xep-0206: Xmpp over bosh, http://bit.ly/xep0206 (verified on February 14, 2012)
24. Augusto, A.B., Correia, M.E.: An xmpp messaging infrastructure for a mobile held security identity wallet of personal and private dynamic identity attributes. In: Proceedings of the XATA 2011 XML: Aplicações e Tecnologias Associadas (2011)
25. Poitner, M.: G&D Secure Flash Solutions. Mobile security card, http://tinyurl.com/SDMSC (verified on February 14, 2012)
26. Maia, L., Correia, M.E.: Java jca/jce programming in android with sd smart cards. In: 7^a Conferencía Ibérica de Sistemas y Tecnologías de Informacións (CISTI 2012), Madrid/ Spain (2012)
27. Bakar, A., Ahmad, A.R., Ismail, R., Manan, J.-L.A.: Trust formation based on subjective logic and pgp web-of-trust for information sharing in mobile ad hoc networks. In: SocialCom 2010, pp. 1004–1009 (2010)
28. Santos, R., Correia, M.E., Antunes, L.: Use of a government issued digital identification card to secure interoperable health information systems. In: The 42nd International Carnahan Conference on Security Technology, ICCST 2008, pp. 1004–1009 (2008)
29. Eastlake, D.: Randomness recommendations for security, http://j.mp/rrsrfc (verified on February 14, 2012)

Care@HOME: A Mobile Monitoring System for Patient Treatment and Blood Pressure Tracking

Mersini Paschou, Efrosini Sourla, George Basagiannis,
Evangelos Sakkopoulos, and Athanasios Tsakalidis

Department of Computer Engineering & Informatics
School of Engineering, University of Patras
Rio Campus, 26500 Patras, Greece
{paschou,sourla,mpasagia,sakkopul,tsak}@ceid.upatras.gr

Abstract. In this paper, we propose an integrated system for mobile monitoring of patient treatment and blood pressure tracking. Care@Home delivers functionality to the patient's smartphone using an intelligent mobile App. It consists of desktop applications and loosely coupled Web Services that allow patient and doctors to interact through either a common Web database or directly through SMS text message mobile notifications. Key features of the proposed solution are (a) the dynamic character of the smartphone App (it is not a static App!), which is possible to receive treatment updates from the doctor remotely and (b) the careful design to deliver alerting on treatment and blood pressure tracking in terms of measurement and instant doctor notification in case of warning levels detection. Care@HOME App stores locally in a smartphone database all measurements taken, beside online Web database storage. In this way, it is possible to minimize data access costs and deliver the measurements during the next doctor visit and only warning level may be transmitted using SMS text messages. Initial evaluation of the prototype has already shown encouraging results.

Keywords: Patient Monitoring, m-Health, Personal Health Records (PHR).

1 Introduction

The spectacular penetration of mobile phones in the technological arena and their transformation into Smartphones has introduced a new field of software applications' development. These applications use limited resources, compared to desktop systems, provided by mobile devices. However, by seizing a number of advantages such as portability, Internet access, location detection services etc. these applications quickly became an integral part of everyday live. Applications related to health care claim a significant market share of mobile applications. These applications mark a new domain, the domain of m-health (mobile health). The number of these applications is great and it is estimated that one out of ten Smartphone users have installed such an application of their device [1]. Projections for 2015 indicate that m-health applications will be used by one out of three users.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 69–83, 2012.

In this paper, we propose a mobile monitoring system for patient Treatment and Blood Pressure Tracking. It is important to have in mind the overview of problems that are related to blood pressure. In fact, increased blood pressure is an important clinical problem as it is quite common, its effects are several times devastating and remains asymptomatic for a long time. Since there is no specific boundary between normal and high blood pressure, the definition of hypertension is arbitrary. This definition takes into account not only the diastolic (low) and systolic (high) pressure but also other factors such as gender, age, and race, body mass, alcohol consumption and heredity. Realizing the serious consequences and devastating effects the chronic presence of elevated levels of blood pressure may have, it is important to emphasize the importance of prevention and systematic observation for the fight against this disease. Recording and quality management of blood pressure measurement in terms of the patient provides a valuable source of information to the physician. Based on this the doctor can monitor the short-term and medium-term course of the disease, to devise a most complete treatment and to assess in detail the performance of a specific medication.

As far as the patient is concerned, engagement in a process of self monitoring increases the level of awareness about the disease and its evolution. The patient becomes responsible for preserving the integrity of his/her health. This procedure becomes a stronger incentive to adopt a course of conduct which includes improved eating habits, increased physical activity and regular intake of treatment to achieve normal blood pressure values. Patients who monitor their blood pressure as a routine are more likely to have normal values [2]. For all the above reasons, we propose an intelligent Smartphone application, together with a back-end complementary system for the physician/clinician, that assist in monitoring treatment and blood pressure issues both in an offline status and while online to directly interconnect the patient with the doctor and vice-versa using typical smartphones that any patient may own.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 presents the proposed system architecture. Section 4 discusses data in Personal Health Records (PHR) software framework. Section 5 presents flowcharts of the available services and related functionality. Section 6 describes the software framework and Section 7 discusses implementation issues for Care@HOME case study. Finally, Section 8 concludes the paper presenting initial evaluation results and future steps.

2 Related Work

Many consumers nowadays take advantage of m-Health applications to improve their lives and assist their health. [3] Benefits of m-Health Applications and Solutions are widely known and accepted. Many existing applications meet the needs of individual specialties in medicine [4] and work in similar ways, whether they are stand-alone applications or they work online. These applications usually have common characteristics; they record critical medical data and communicate with other applications in an effort to solve a health issue. [5] Data related to health records are often sent to servers storing personal health record services or directly to physicians. [6]

A Personal Health Record (PHR) is a folder where health information is organized and accessed by the person the record concerns. This record can be in printed form, or can be implemented locally in an electronic device e.g. a PC or mobile phone or finally, it can be implemented on the web as a web application. In this paper, our interest is focused mostly on the third category of implementation, which is the web, which can be accessed using a mobile device. The major contribution of this work is the design of a personalized, integrated, and collaborative care system for self-monitoring and tracking of Blood Pressure.

In [7] early diagnosis of hypertension and other chronic diseases is attempted, with a system has three main parts: a wrist Blood Pressure measurement unit, a server unit and a terminal unit. Blood Pressure is detected and the data acquired by sensors intelligently. The data is then transmitted to the remote server unit located at Community Healthcare Centers/Points by using Short Messaging Service (SMS), and notification information is sent to the terminal unit to inform users if patient's Blood Pressure is abnormal. A personal diabetes monitoring system is proposed in [8], which integrates wearable sensors, 3G mobile phone, smart home technologies and Google Health to facilitate the management of chronic disease - diabetes.

A small device for noninvasive continuous measurement of blood pressure is developed in [9]. The device used is a small battery powered embedded system which measures blood pressure during a long time period and sends this information remotely to operator's stations. In this station the analysis of receiving video signal is made for several types of usage. The video signal is transferred by wireless connection to desktop station where it is analyzed.

In our work, we enhance blood pressure monitoring with a system that combines in one -publicly available- mobile application monitoring, treatment as well as communication. It enables the user to record his/her measurements at any given time using a mobile application and allows for direct communication with a physician in cases of emergency. Moreover, the physician can update the treatment or medication of the patient when necessary and any changes are sent to the mobile application. Data is stored locally as well as online and are available for monitoring.

3 Care@Home System Architecture

In this work we present an overview of the integrated system for blood pressure monitoring. The system is called Care@HOME and consists of two separate applications that interact with a common database. On the one hand we have a desktop application which is used by the physician and on the other hand we have a mobile application used by the patient. In our design we use Microsoft HealthVault as Personal Health Record (PHR) management system. With this tool we aim at bringing together the patient and everyone involved in his therapeutic procedure (physicians, relatives etc), towards a process that allows everybody to work together and refine goals for the person whose health is monitored.

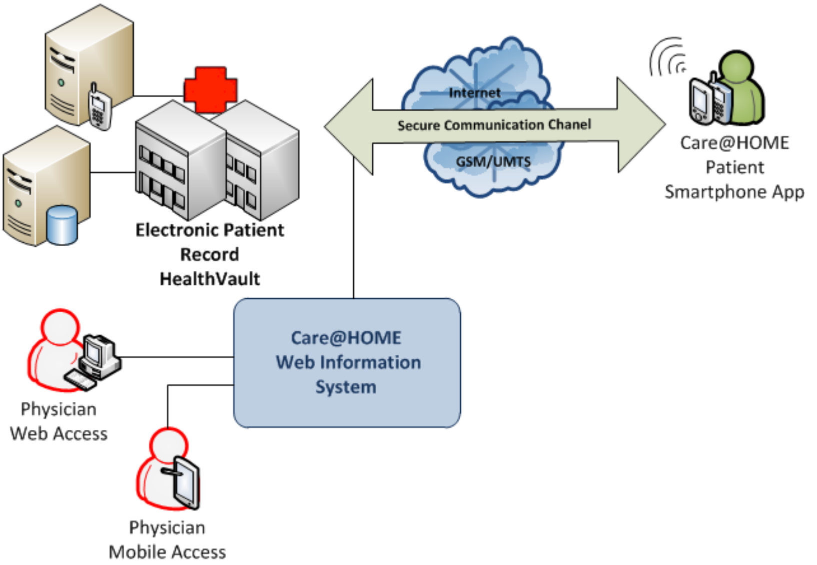


Fig. 1. Care@HOME System Architecture

The physician can access Care@Home system using web or mobile access in order to determine quality parameters related to the fluctuation of blood pressure levels or modify therapeutic action of the user-patient. The patient can use the Care@HOME Smartphone app which receives the settings set by the physician for the specific user and manages the data that the user enters, saving them both locally and online. (Fig. 1)

4 Data in Personal Health Record (PHR)

The characteristic that sets PHR apart from the widely used Electronic Medical Records (EMRs) is the fact that it is organized and controlled by the user-patient himself. EMRs are monitored by organizations such as hospitals and contain information recorded by health clinicians and / or information related to hospital costs which are addressed to insurance agencies.

The purpose of the PHR is to provide a comprehensive medical history, available and accessible online. The information included may be data entered by the user himself, laboratory results, data extracted from wireless measuring devices etc. PHR as a service is provided mainly by private companies. A detailed list of PHR services is available online [10].

4.1 Advantages Personal Health Record (PHR) and Data

Usage of PHR provides the user with many advantages, including storage of health information in a safe place and users can easily access to retrieve information about their medical history. This information may be shared, using authorization, with a personal physician, health professionals or even family members [11], [12]. Moreover, having access to the Internet the user can fill in and upgrade his/her personnel file with the current state of health. Through the usage of PHR it is easier to organize medical information and plan future events such as vaccinations and preventive examinations (e.g. scheduling mammograms). Health information is no longer fragmented between different health professionals, hospitals and insurance agencies. The user becomes more active in his health care plan by collecting, organizing and archiving the related information. More specifically, the features related to HealthVault are:

- Storage and handling of health information
- Creation of emergency profile
- Ability to add measurements and data, using compatible devices
- Review and analysis of the stored information using the provided online tools
- Sharing part or all of the stored information with doctors or family members [14].

In urgent cases, a personal health file can be a valuable resource for the physician, as the patient's history, chronic diseases, allergies, medication of the patient and any other critical information can be accessed directly. Adopting a collective policy for using PHR, in addition to the benefits mentioned above, is estimated to be of significant benefit for insurance funds while reducing the cost of health expenditure. Related studies estimate an expected cost reduction of amounts ranging from 13 to 21 billion [15, 13].

The personal health file may contain a fairly wide range of information related directly or indirectly to the health of the user. More specifically:

- Personal information i.e. name, date of birth and current address
- Names and phone numbers of relatives or people of the owner's friendly environment that can be contacted in case of emergency
- Names, addresses and telephone numbers of physicians
- Info related to individual health insurance
- Current medication (if any) and respective dosages
- Known allergies to foods, drugs and other substances
- Important events, dates and hereditary diseases, involving the family history
- History of the most important diseases encountered by the user in his past
- Results of medical examinations, important medical tests, dental history as well as vaccination history
- Recent medical diagnostics, summary of visits to family physician or another specialist
- Information related to physical activity, exercise program, dietary restrictions and record of medications that do not require a prescription (over the counter - OTC) and / or alternative therapeutic approaches.

All these capabilities render HealthVault to be a valuable tool of import and management of health related information. A major limitation should be stressed: using PHR is available only to residents of the United States of America, due to legal obstacles. It is the company's intention to expand its use in other countries, provided that the relevant legal restrictions will be eliminated.

Unlike using it, development of software which supports or uses HealthVault is possible outside the United States as well. For this purpose Microsoft HealthVault Pre-Production Environment (PPE) is available. The PPE is a web server platform that simulates the HealthVault website except that it does not provide access to real user data. It has been designed specifically to support the development of related applications, so that developers can test and evaluate their software. The Blood Pressure Tracker uses such an account, created in PPE.

5 Flowchart of Services and Functionality

The available services provided by Care@HOME Web Information System are presented below in detail.

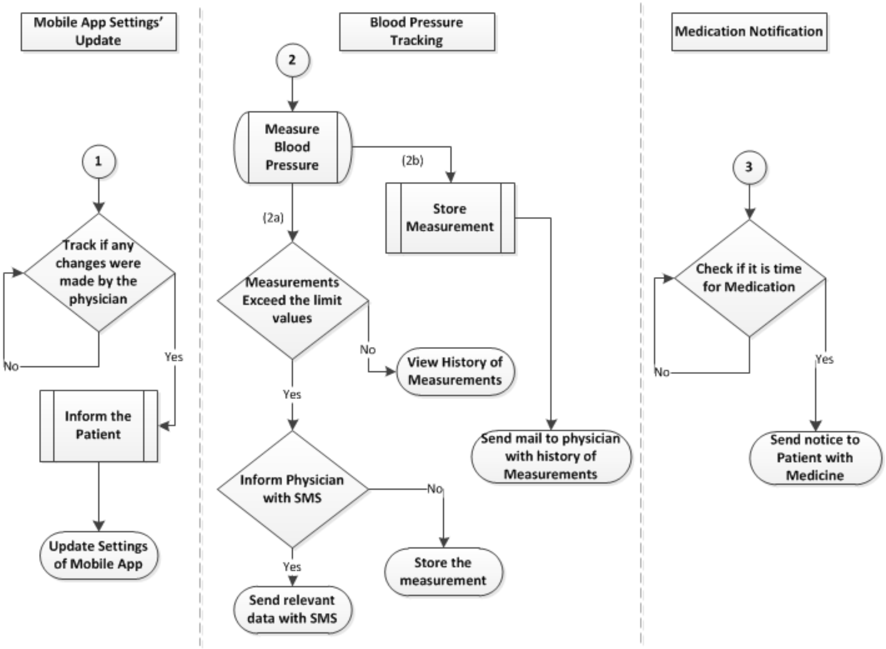

Mobile App Settings' Update
Blood Pressure Tracking
Medication Notification
2
Measure Blood Pressure
(2b)
Store Measurement
(2a)
1
Track if any changes were made by the physician
No
Yes
Inform the Patient
Update Settings of Mobile App
Measurements Exceed the limit values
No
View History of Measurements
Yes
Inform Physician with SMS
No
Yes
Send relevant data with SMS
Store the measurement
Send mail to physician with history of Measurements
3
Check if it is time for Medication
No
Yes
Send notice to Patient with Medicine

Fig. 2. Care@HOME Flowchart for provided services

1. The integrated system will track any changes made by the physician and will inform the patient and will automatically update the relevant settings of the mobile application.

2. Through the mobile application the user can add new blood pressure measurements, which are stored locally on the device and online at HealthVault.

(a) The application checks the incoming measurements each time and compares them to the limits set by the physician. If the measurements exceed the limit values, the application informs the user that a text message (sms) should be send to the physician. The patient approves it and the application undertakes sending the relevant data via sms. The user can look at previous measurements and see illustrations of these measurements with full representation of all data. It is also possible to view the history of measurements through charts.

(b) The user may send email to the physician, with the values he/she has registered in the past and choose the period to which they relate.

3. The application informs the patient for taking the medicines corresponding to his/her treatment. This is achieved with a notification, which is triggered regardless whether the application closed or not. A notice is sent to the patient, which includes the name of the medicine he/she should take at that moment, the appropriate dosage and other helpful information e.g. the shape and color of the formulation. The last characteristic is particularly useful in case a patient receives more than one treatment.

Additional Services

- An update via a twitter client on the latest news related to hypertension published on twitter is also possible.
- Additionally, contact information of the physician (phone, email) may be altered by the user of the mobile application.
- Moreover, the personal id can be changed and the user has the ability to change the language of the application (Greek or English).
- In the settings, there is support for the activation of operation of simple voice instructions in languages other than the two mentioned above, e.g. Spanish, French etc. This function is served through Bing Translate.

6 Software Framework

Multi-module software architecture is developed (Fig. 3) for the desktop part in Visual Studio 2010 with the programming language C# and the database is stored in Microsoft SQL server 2008. The mobile application is implemented in Windows Phone 7, the new platform for Smartphone of Microsoft, in specific version 7.5 Mango, which is an upgraded version of the initial.

Smartphone App GUI Module allows the patient to input or retrieve data by interacting with it. The dedicated applications include: 1) Adding and modifying blood pressure measurements. 2) Viewing previous measurements and illustrations of these measurements with full representation of all data. 3) Viewing history of measurements through charts. 4) Sending email to the physician, with values registered in the past. 5) Updating on latest news related to hypertension through a twitter client. 6) Modification of contact information of the physician. 7) Alteration of the language of the application (Greek or English).

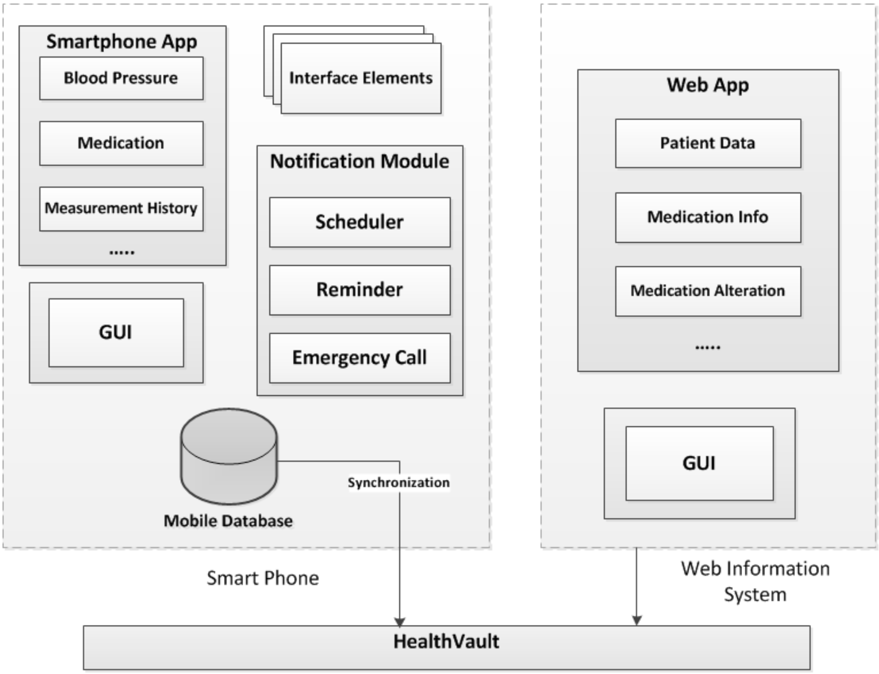


Fig. 3. Care@HOME Software Framework

Notification Module is responsible for reminders to the user-patient and the physician or any other person indicated as contact person. In emergency situations, that is when blood pressure is out of normal values; notification module will alert the user and contact the physician, who will then take some action, if necessary.

Database Module enables storage of crucial information of the patient locally and allows for synchronization with HealthVault whenever it is possible. Essentially it is needed for the physician to be able to retrieve health data of the patient when necessary.

Web Application GUI Module in .NET enables the physician to input or retrieve data by interacting with it. The applications include: 1) Patient information input or alteration. 2) Medication information for a specific patient. 3) Medication modification for a specific patient.

Web Service is deployed to enable communication between the Smartphone application and HealthVault It also allows for exchange of data amongst HealthVault and the web application used by the physician.

Microsoft HealthVault is a web based PHR system. It is a free online tool that stores health information records in a central location. With the help of special health tools, information is accessible and usable by the user.

7 Care@HOME Case Study

As mentioned above the proposed Care@HOME system for blood pressure monitoring and patient treatment consists of two applications loosely interconnected to a central database and which interact. The desktop application is operated by the physician and the mobile application is used by the patient. All case study software has been developed using Microsoft .NET technologies. The smartphone application is delivered for Windows Phone 7.5 OS. The software development platform is based on existing tools like MS Visual Studio, Expression Blend and Silverlight. In this way our developers who have experience with these tools can develop applications for Windows Phone and Desktop OS without having to spend extra time on studying the tools. Please notice, that the generality of the proposed architecture allows us to extend already the care@HOME App to additional platforms such as Android platforms for smartphones and tablets.

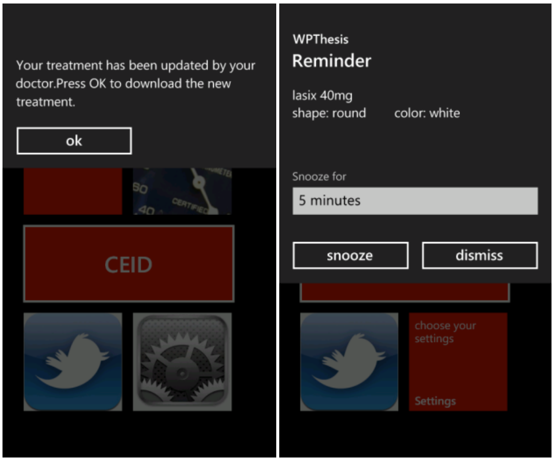
7.1 Smartphone App

One part of the Care@HOME blood pressure tracking system is the mobile application. The mobile operator is the user-patient. In launch, it is connected via a WCF service to the central database. It detects if any changes of settings have been made by the physician. If changes have occurred, it informs the user and proceeds to update of the local settings. The user through the application can add new blood pressure measurements, which are stored locally on the device and online at HealthVault. The user can refer to the history of previous measurements and see illustrations of these measurements with full representation of all data. It is also possible to see representation of the history of measurements through charts.

The application user has the ability to send email to the doctor with the measurement values which have been stored in the past and choose the period to which they relate. Moreover, he/she can be updated via a twitter client for the latest news on twitter related to hypertension. Having access to application settings, the user can reenter the contact information of the physician (phone, email), change his/her personal id and has the ability to change the language of the application. The application is bilingual and fully supports both Greek and English. Additionally, in settings the function of simple voice instructions can be activated, in languages other than the two mentioned above, e.g. Spanish, French etc. This function is served through Bing Translate.

In addition to the actions of the user, the application is enabled to inform of any actions he/she must take. Each time there are incoming measurements, the application compares them with the limits set by the physician. If the measurements exceed the limit values, the application informs the user that a text message (SMS) should be sent to the doctor. The patient gives approval for this action and the application undertakes sending the appropriate data via SMS. Last but not least, the application must notify the patient when it's time for receiving medications that correspond to the treatment. This is done in the form of notification, and regardless whether the application is on the foreground or not (closed application) the patient receives a notice of the medicine he/she should take, the dosage and other helpful information, e.g. the shape and color of the formulation. The last characteristic is particularly useful in case the patient receives more than one treatment.

Before the user gains control of the application, the central base is accessed to explore changes made from the physician. If changes have been made the application informs the user. Ultimately the patient will receive a notification for any medication on the mobile screen, at the time set, along with the corresponding alarm. (Fig. 4)


Your treatment has been updated by your doctor.Press OK to download the new treatment.

WPThesis
Reminder
lasix 40mg
shape: round color: white
Snooze for
5 minutes

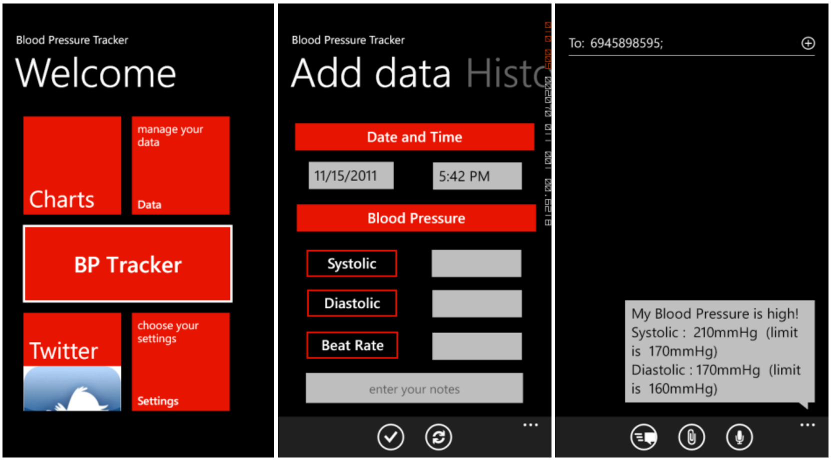
Fig. 4. (a) Message for Updated treatment available to download, (b) Reminders for treatment-medication/drug

Once all processes have been configured in the background, the first screen that appears consists of a main menu with four controls that belong to the hub tile category. These are dynamic tiles that are enabled to move. They give the sense of movement upward, downward and rotate while displaying information. (Fig. 5-a) The options that will be seen by the user, to move to the relevant parts of the application are the following:

- Data: the option related to addition, review and distribution of medical information
- Charts: the part of the application that includes visualization of medical data through graphs
- Twitter: provides access to twitter client
- Settings: provides access to settings.

The information management page is called Data and is implemented by a Pivot control. A pivot control in Windows Phone is an easy and accessible way of presenting information, with changes between forms.

Through the respective controls date and time of measurement can be selected and values of systolic and diastolic pressure can be registered. In addition, pulse (heartbeats per minute) is recorded and optionally auxiliary notes. Notes can be associated with physical and / or mental state of the user during the measurement and can be general information to complement the profile of each record. (Fig. 5-b) With the submission of the measurement by the user, a record is saved both locally on the device and online at HealthVault. Along with the inclusion of the new measurement, a value control of the systolic and diastolic pressure is made. If the value of any of them exceeds the relevant limits set by the doctor then the application informs the user that they should send a text message informing the attending physician.



11/15/2011 5:42 PM

To: 6945898595;

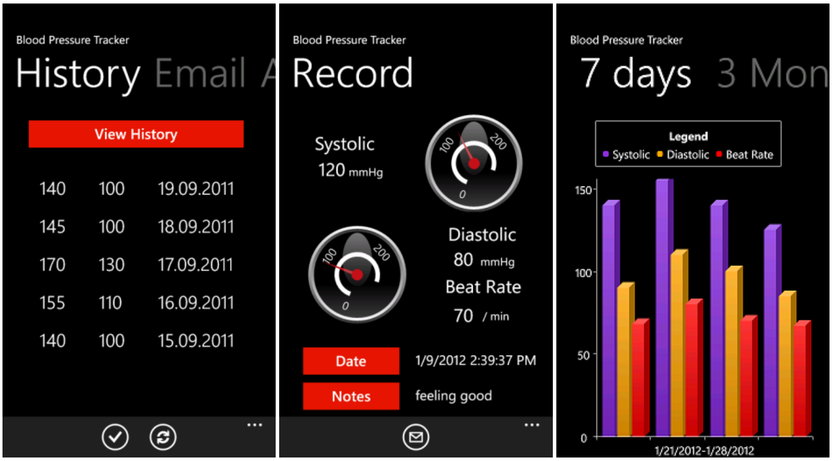
My Blood Pressure is high!
Systolic : 210mmHg (limit is 170mmHg)
Diastolic : 170mmHg (limit is 160mmHg)

Fig. 5. (a) Main Panel Options, (b) Record blood pressure values, (c) SMS to doctor about High blood pressure

Upon acceptance by the user, a text message with the relevant information is formed. Sending the written message is indicated by the user; the application prepares the message but does not send it automatically. (Fig. 5-c) The user will choose whether to send or revoke it. The user can also browse the list by dragging it up or down and choose a measurement which he will view in detail. In the detailed page, systolic and diastolic pressure are depicted, using the gauge control, while other information of the measurement are provided, such as pulse, date and time of recording and notes recorded by the user. (Fig. 6.) Moreover, he/she can view, in graphical form, representations of the measurements that have been registered. (Fig. 6-a) Graphs can refer to all measurements stored or report only specific time intervals selected by the user.

Social Networking Functionality: The user can navigate to an initial list that includes the last fifty tweets on hypertension. He/she can choose to see more information about one of them. If the specific tweet includes a hyperlink, the user can navigate to the relevant web page to view the desired information. (Fig. 7 -a)

Despite the fact that the twitter client does not provide scientific information on hypertension, it has been added for recreational purposes, giving the user the opportunity to see what is discussed on a popular social network in the world on the health issue he is interested in learning more about.


Blood Pressure Tracker

History Email A

View History

140	100	19.09.2011
145	100	18.09.2011
170	130	17.09.2011
155	110	16.09.2011
140	100	15.09.2011

Blood Pressure Tracker

Record

Systolic 120 mmHg

Diastolic 80 mmHg

Beat Rate 70 / min

Date 1/9/2012 2:39:37 PM

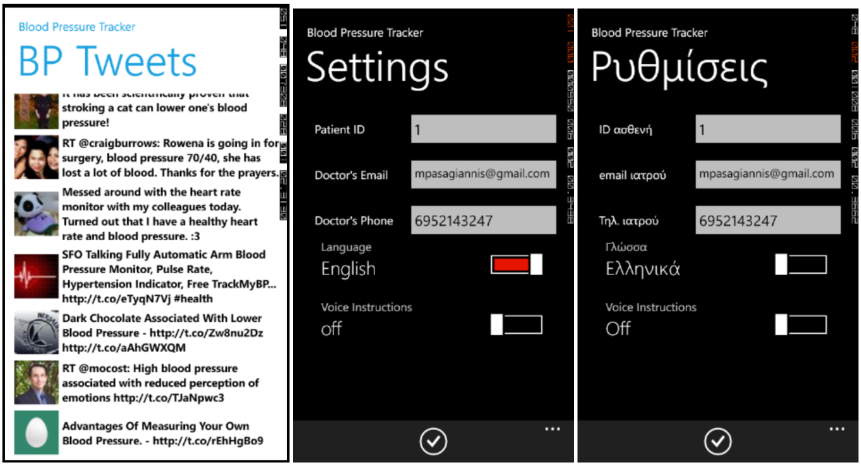
Notes feeling good

Blood Pressure Tracker

7 days 3 Mon

Fig. 6. (a) History log of recorded values, (b) Detail View of a recorded measurement using gauge control display (c) charts on systolic, diastolic and pulses of a week's recorded values

Localization: the process aims at adapting an application to the needs of people living in different geographic regions around the globe. Localization does not only include adapting the language of an application but other factors as well, such as the emergence date, time, currency, etc. depending on the circumstances of the area being targeted. The Care@HOME application supports English (en-US) and Greek (el-GR).


Blood Pressure Tracker

BP Tweets

It has been scientifically proven that stroking a cat can lower one's blood pressure!

RT @craigburrows: Rowena is going in for surgery, blood pressure 70/40, she has lost a lot of blood. Thanks for the prayers.

Messed around with the heart rate monitor with my colleagues today. Turned out that I have a healthy heart rate and blood pressure. :3

SFO Talking Fully Automatic Arm Blood Pressure Monitor, Pulse Rate, Hypertension Indicator, Free TrackMyBP... http://t.co/eTyqN7Vj #health

Dark Chocolate Associated With Lower Blood Pressure - http://t.co/Zw8nu2Dz http://t.co/aAhGWXQM

RT @mocost: High blood pressure associated with reduced perception of emotions http://t.co/TJaNpwc3

Advantages Of Measuring Your Own Blood Pressure. - http://t.co/rEhHgBo9

Blood Pressure Tracker

Settings

Patient ID 1

Doctor's Email mpasagiannis@gmail.com

Doctor's Phone 6952143247

Language
English

Voice Instructions
off

Blood Pressure Tracker

Ρυθμίσεις

ID ασθενή 1

email ιατρού mpasagiannis@gmail.com

Τηλ. ιατρού 6952143247

Γλώσσα
Ελληνικά

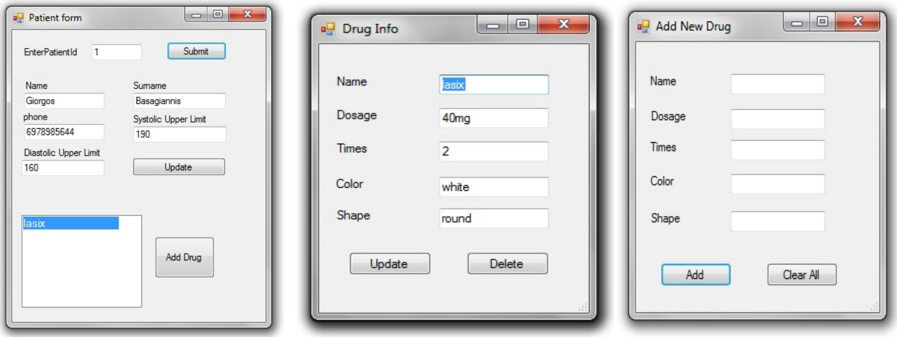
Voice Instructions
Off

Fig. 7. (a) Blood pressure tweet information, (b-c) settings page (phone number of doctor, language etc) in Greek and in English to show the localization effects

7.2 Desktop Software

The desktop software includes forms with an oversight of patient information and medication received. The physician can modify the information of the patient by changing the relevant inputs. The name, telephone, and the upper limits of systolic and diastolic pressure can be altered. After entering the desired values in the relevant fields, the relevant event can be triggered by pressing the update button and the contents of fields are sent to the database.

In the case that the patient selected already receives a treatment, then the treatment is presented to a respective listbox. Choosing a drug from the listbox opens a new form, called Drug Info, which provides detailed information about this medicine and allows renewal of the fields of medicine or removal from the treatment regimen. Next to the listbox there is a button titled "Add Drug" which opens a third form, the "Add New Drug" form, which is used by the physician to add a new drug regimen to the patient. (Fig. 8)



EnterPatientId: 1

Name: Giorgos

Surname: Basagiannis

phone: 6978985644

Systolic Upper Limit: 190

Diastolic Upper Limit: 160

lasix

Name: lasix

Dosage: 40mg

Times: 2

Color: white

Shape: round

Name:

Dosage:

Times:

Color:

Shape:

Fig. 8. Doctor's Desktop Client

The physician can modify the data e.g. the drug dose or the number of daily reception. The modification process is achieved by pressing the Update button, which triggers the relative event.

7.3 Web Services and HealthVault – SmartPhone Integration

A prerequisite for developing applications in windows phone 7 that will use HealthVault is installation of the library Microsoft HealthVault Windows Phone which is available in http://healthvaultwp7.codeplex.com/. The library enables encrypted communication among the application and the server of HealthVault. The respective SDK provides supplementary tools for interaction with the platform, instructions for the programmer as well as some standard applications.

Each application is represented by a unique ID, which is received through HealthVault Application Configuration Center. To access it, an account in PPE is necessary. After introducing our data we can navigate in the Application Configuration Center. The next step is the creation of a new application and selection of its name and type (SODA). It is important is to define the type of data to be sent by the application. At the main menu we select Soda app rules and then we choose the data type "Blood Pressure Measurement". We store the settings and note the ID of the application.

8 Evaluation, Conclusions and Future Steps

This paper presents a personal blood pressure monitoring system Care@HOME which integrates 3G mobile phones, web applications, smart technologies and Microsoft HealthVault to facilitate the management of health related data. We have shown how typical smartphones may assist patient monitoring at home. Doctors' reviews and patients' feedback on the Care@HOME smartphone application for treatment and blood pressure monitoring have shown encouraging results during our evaluation with University of Patras Hospital staff. They have particularly mentioned that it has a user friendly UI for the application – system especially in terms of reminding the user to perform his measurement on blood pressure and to receive his/her medication. Review also show that the proposed App can be particularly valuable as it enables the patient to share monitoring information with the doctor "on-time" using either SMS or HTTP in a simple manner for both ends. Feedback was also positive for the social media interconnection through twitter preloaded access.

Creation of a blood pressure tracker system aims to provide a comprehensive tool for assisting hypertension issues. The doctor, additionally to the initial determination of the regimen, is enabled to redefine treatment by evaluating the data received and to forward it to the patient directly. On the other hand the patient, through the systematic measurement of the values of blood pressure and the ability to review organized and classified information, that he entered himself, becomes more active and responsible in matters related to personal health. Moreover, the patient feels more secure because he/she knows that the physician is aware of course of the disease and is able to intervene directly and attending the clinic is not a prerequisite.

Future steps include the interconnection of the alerting mechanisms to a decision support system (DSS) that could additionally decide in an autonomous way whether the patient should be alerted or automatically send crucial information to the doctors in case of emergency. Additionally, it is within our intention to deliver and distribute the initial version of Care@HOME to additional smartphone platforms for even more markets and stores to be covered.

Acknowledgements. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. Global Mobile Health Market Report 2010-2015 (March 24, 2012), http://www.research2guidance.com/500m-people-will-be-using-healthcare-mobile-applications-in-2015/
2. Use of The Carrot.com‘s Hypertension Tracker Soars 30% During Heart Month (March 24, 2012), http://www.reuters.com/article/

3. Lytras, M., Sakkopoulos, E., de Pablos, P.O.: Semantic Web and Knowledge Management for the health domain: state of the art and challenges for the Seventh Framework Programme (FP7) of the European Union (2007–2013). International Journal of Technology and Management (IJTM) 47(1/2/3), 239–249 (2009)
4. Gkintzou, V., Papablasopoulou, T., Syrimpeis, V., Sourla, E., Tzimas, G., Tsakalidis, A.: A Web and Smart Phone System for Tibia Open Fractures. In: Cruz-Cunha, M.M., Varajão, J., Powell, P., Martinho, R. (eds.) CENTERIS 2011, Part III. CCIS, vol. 221, pp. 413–422. Springer, Heidelberg (2011)
5. Paschou, M., Zorba, I., Sakkopoulos, E., Tsakalidis, A.: APPification of Hospital Health-Care and Management. In: International Conference on Information Communication Technologies in Health (ICICTH), Samos Island, Greece, July 12-14 (2012)
6. Paschou, M., Papadimitriou, C., Sakkopoulos, E., Tsakalidis, A.: Personnel Rostering System in Health Care Units using Mobile Technologies. In: 5th World Summit on the Knowledge Society (WSKS), Rome, Italy, pp. 20–22 (June 2012)
7. Jiang, J., Yan, Z., Shi, J., Kandachar, P.: Design of Wireless Mobile Monitoring of Blood Pressure for Underserved in China by Using Short Messaging Service. In: Proceedings of the 5th International Conference on Information Technology and Application in Biomedicine, Shenzhen, China, May 30-31 (2008)
8. Zhou, F., Yang, H.-I., Álamo, J.M.R., Wong, J.S., Chang, C.K.: Mobile Personal Health Care System for Patients with Diabetes
9. Krejcar, O., Janckulik1, D., Motalova, L., Frischer, R.: Architecture of Mobile and Desktop Stations for Noninvasive Continuous Blood Pressure Measurement. In: Proceedings of WC 2009. IFMBE, vol. 25, pp. 137–140 (2009)
10. A detailed list of PHR services (March 24, 2012), `http://www.myphr.com/`
11. Frequently Asked Questions on Personal Health Record (PHR) (March 24, 2012), `http://www.myphr.com/resources/faqs.aspx`
12. Managing Your Health Information Online (March 24, 2012), `http://www.medicare.gov/navigation/manage-your-health/personal-health-records/personal-health-records-overview.aspx`
13. A Cost-Benefit Model for PHRs (March 24, 2012), `http://journal.ahima.org/2008/11/17/a-cost-benefit-model-for-phrs/`
14. Microsoft HealthVault (March 24, 2012), `http://www.microsoft.com/en-us/healthvault/`
15. Kaelber, D., Pan, E.C.: The value of personal health record (PHR) systems. Centre for Information Technology Leadership (CITL), Partners HealthCare System, Boston, MA, USA

An Integrative Clustering Approach Combining Particle Swarm Optimization and Formal Concept Analysis

Anna Hristoskova[1], Veselka Boeva[2], and Elena Tsiporkova[3]

[1] Department of Information Technology,
Ghent University - IBBT, 9050 Ghent, Belgium
anna.hristoskova@intec.UGent.be

[2] Department of Computer Systems and Technology
Technical University of Sofia-branch Plovdiv 4000 Plovdiv, Bulgaria
vboeva@tu-plovdiv.bg

[3] ICT & Software Engineering Group, Sirris, 1030 Brussels, Belgium
elena.tsiporkova@sirris.be

Abstract. In this article we propose an integrative clustering approach for analysis of gene expression data across multiple experiments, based on Particle Swarm Optimization (PSO) and Formal Concept Analysis (FCA). In the proposed algorithm, the available microarray experiments are initially divided into groups of related datasets with respect to a predefined criterion. Subsequently, a hybrid clustering algorithm, based on PSO and k-means clustering, is applied to each group of experiments separately. This produces a list of different clustering solutions, one per each group. These clustering solutions are pooled together and further analyzed by employing FCA which allows to extract valuable insights from the data and generate a gene partition over the whole set of experiments. The performance of the proposed clustering algorithm is evaluated on time series expression data obtained from a study examining the global cell-cycle control of gene expression in fission yeast *Schizosaccharomyces pombe*. The obtained experimental results demonstrate that the proposed integrative algorithm allows to generate a unique and robust gene partition over several different microarray datasets.

Keywords: data clustering, k-means, particle swarm optimization, formal concept analysis, integration analysis, gene expression data.

1 Introduction

DNA microarray technology offers the ability to screen the expression levels of thousands of genes in parallel under different experimental conditions or their evolution in discrete time points. All these measurements contain information on many different aspects of gene regulation and function, ranging from understanding the global cell-cycle control of microorganisms [20], to cancer in humans [1,10]. Gene clustering is one of the most frequently used analysis methods for gene expression data. Clustering algorithms are used to divide genes into

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 84–98, 2012.

groups according to the degree of their expression similarity. Such a grouping may suggest that the respective genes are correlated and/or co-regulated, and moreover that the genes could possibly share a common biological role.

The combination of data from multiple microarray studies addressing a similar biological question is gaining high importance in the recent years [6,7,24] due to the ever increasing number and complexity of the available gene expression datasets. In general, it is expected that the integration and evaluation of multiple datasets yields more reliable and robust results since these results are based on a larger number of samples and the effects of individual study-specific biases are diminished. In [5], we proposed a hybrid algorithm combining k-means and Particle Swarm Optimization (PSO) clustering algorithms in order to derive a gene clustering solution from a set of independent, but biologically related, microarray datasets. It was demonstrated that this hybrid algorithm produces good quality clustering solution, which is representative for the whole experimental compendium and at the same time adequately reflects the specific characteristics of the individual experiments.

In this work, we propose an integrative clustering method that combines PSO and Formal Concept Analysis (FCA) [9] in order to cluster datasets generated in multiple-experiment settings. In contrast to the hybrid clustering algorithm introduced in [5], where PSO-based clustering is applied to the entire set of experiments in order to produce the final clustering solution, the algorithm proposed in this paper initially divides the available microarray experiments into groups of related (similar) datasets with respect to a predefined criterion. The rationale behind this is that if experiments are closely related to one another, then these experiments may produce more accurate and robust clustering solution. Thus PSO-based clustering is applied to each group of experiments separately. This produces a list of different clustering solutions, one per each group. Next these solutions are pooled together and further analyzed by employing FCA which allows to extract valuable insights from the data and further generate a gene partition over the whole experimental compendium. FCA produces a concept lattice where each concept represents a subset of genes that belong to a number of clusters. The concepts compose the final disjoint clustering partition.

A detailed overview of several PSO-based clustering approaches is presented in [5]. The FCA or *concept lattice approach* has been applied for extracting local patterns from microarray data [2,3] or for performing microarray data comparison [8,18]. For example, the FCA method proposed in [8] builds a concept lattice from the experimental data together with additional biological information. Each vertex of the lattice corresponds to a subset of genes that are grouped together according to their expression values and some biological information related to the gene function. It is assumed that the lattice structure of the gene sets might reflect biological relationships in the dataset. In [13], a FCA-based method is proposed for extracting groups or classes of co-expressed genes. A concept lattice is constructed where each concept represents a set of co-expressed genes in a number of situations. A serious drawback of the method is the fact that the expression matrix is transformed into a binary table (the input for the FCA step) which may lead to possible introduction of biases or substantial information loss.

The remainder of this paper is structured as follows: Section 2 briefly describes the basic principles of k-means, PSO, hybrid clustering and FCA, and subsequently introduces our integrative clustering approach. The dataset and the applied experimental setup are outlined in Section 3, followed by analysis and discussion of the clustering results in Section 4. Finally, the main conclusions are drawn in Section 5.

2 Clustering Methods

2.1 K-means Clustering Algorithm

The *k-means algorithm* [15] is one of the most widely used techniques for clustering. It starts by initializing the k cluster centers, where k is preliminarily determined. Subsequently, each object (input vector) of the dataset is assigned to the cluster whose center is the nearest. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the objects and the update of the cluster centers is repeated until there is no more change in the value of any of the cluster centers.

2.2 Particle Swarm Optimization

Particle swarm optimization (PSO) is an evolutionary computation method introduced in [14]. In order to find an optimal or near-optimal solution to the problem, PSO updates the current generation of particles (each particle is a candidate solution to the problem) using the information on the best solution obtained by each particle and the entire population. Each particle is treated as a point in an n-dimensional space. The i-th particle is initialized with random positions $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and velocities $V_i = (v_{i1}, v_{i2}, \ldots, v_{in})$ at time point $t = 0$. The performance of each particle is measured according to a predefined fitness function, which uses the particle's positional coordinates as input values. Positions and velocities are adjusted, and the function is evaluated with the new coordinates at each time-step. The basic update equations for the d-th dimension of the i-th particle in PSO may be given as

$$v_{id}(t+1) = w \cdot v_{id}(t) + c_1 \cdot \varphi_1 \cdot (p_{id} - x_{id}(t)) + c_2 \cdot \varphi_2 \cdot (p_{gd} - x_{id}(t)) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1). \quad (2)$$

The variables φ_1 and φ_2 are uniformly generated random numbers in the range $[0, 1]$, c_1 and c_2 are called acceleration constants whereas w is called inertia weight [21]. $P_g = (p_{g1}, p_{g2}, \ldots, p_{gn})$ is the best particle position found so far within the population and $P_i = (p_{i1}, p_{i2}, \ldots, p_{in})$ is the best position discovered so far by the corresponding particle. The first part of equation (1) represents the *inertia* of the previous velocity, the second part is the *cognition part* and it tells us about the personal experience of the particle, the third part represents the cooperation among particles and is therefore named the *social component*. Acceleration numbers c_1, c_2 and inertia weight w are predefined by the user. It was shown in [21] that when w is in the range $[0.9, 1.2]$, PSO will have the best chance to find the global optimum within a reasonable number of iterations. Furthermore, $w = 0.72$ and $c_1 = c_2 = 1.49$ were found in [17] to ensure good convergence.

Notice that in the multi-experimental context considered in Section 2.3 the *cognition part* representing the personal opinion of the particle is based on its own source of information (dataset), while in the classical one dataset setup the *cognition part* is derived from a common (for all particles) information source. This may also have a reflection on the *social part*, since information contained in different sources may have different representations and may need to be preprocessed before the collaboration of the particles.

2.3 Hybrid PSO-Based Approach for Clustering Data Compendiums

We have proposed in [5] a hybrid algorithm combining k-means and PSO for deriving a clustering result from multiple microarray datasets. This algorithm will be used to produce a clustering result from a group of related microarray experiments in Section 2.5. The main idea of the algorithm and its consecutive steps are presented below.

Let us consider a group of n different microarray datasets $M_1, M_2, \ldots, M_n$. Each dataset is supposed to contain the gene expression levels of m genes in n_i different experimental conditions or time points. In this context, each matrix i can be used to generate k cluster centers, which are considered to represent a particle, *i.e.* the particle is treated as a set of points in an n_i-dimensional space. The final (optimal) clustering solution will be found by updating the particles using the information on the best clustering solution obtained by each data matrix and the entire set of matrices.

Assume that the i-th particle is initialized with a set of k cluster centers[1] $C_i = \{C_1^i, C_2^i, \ldots, C_k^i\}$ and a set of velocity vectors $V_i = \{V_1^i, V_2^i, \ldots, V_k^i\}$[2] using gene expression matrix M_i. Thus each cluster center is a vector $C_j^i = (c_{j1}^i, c_{j2}^i, \ldots, c_{jn_i}^i)$ and each velocity vector is a vector $V_j^i = (v_{j1}^i, v_{j2}^i, \ldots, v_{jn_i}^i)$, *i.e.* each particle i is a matrix (or a set of points) in the $k \times n_i$ dimensional space.

Next, assume that $P_g = \{P_{g1}, P_{g2}, \ldots, P_{gk}\}$ is a set of cluster centers in an n_g-dimensional space representing the best clustering solution found so far within

[1] The number of clusters k, is initially identified by analyzing the quality of the obtained clustering solutions generated on the involved datasets for a range of different numbers of clusters.

[2] The velocity vectors are initialized by zeros.

the set of matrices and $P_i = \{P_1^i, P_2^i, \ldots, P_k^i\}$ is the set of centroids of the best solution discovered so far by the corresponding matrix. The update equation for the d-th dimension of the j-the velocity vector of the i-th particle is defined as follows

$$v_{jd}^i(t+1) = w \cdot v_{jd}^i(t) + c_1 \cdot \varphi_1 \cdot (p_{jd}^i - c_{jd}^i(t)) + c_2 \cdot \varphi_2 \cdot g(t), \tag{3}$$

where $i = 1, \ldots, n$; $j = 1, \ldots, k$; $d = 1, \ldots, n_i$ and

$$g(t) = \begin{cases} p_{gd} - c_{jd}^i(t), & \text{if } n_g \geq n_i \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

Note that the *cognition part* in the above equation has a modified interpretation. Namely, it represents the private 'thinking' (opinion) of the particle based on its own source of information (dataset). Due to this the *social part* (see equation (4)) differs from that in equation (1), since each particle matrix has a different number of columns (n_i) due to different number of experiment points in each dataset.

The clustering algorithm combining PSO and k-means can be summarized as follows:

```
1. Initialize each particle with $k$ cluster centers obtained as a result of applying
   the k-means algorithm to the corresponding data matrix.
2. Initialize the personal best clustering solution of each matrix with the corresponding
   clustering solution found in Step 1.
3. for iteration = 1 to max-iteration do
   (a) for $i = 1$ to $n$ do (i.e. for all datasets)
         i. for $j = 1$ to $m$ do (i.e. for all genes in the current dataset)
              A. Calculate the distance of gene $g_j$ with all cluster centers.
              B. Assign $g_j$ to the cluster that has the nearest center to $g_j$.
        ii. end for
       iii. Calculate the fitness function for the clustering solution $C_i$.
        iv. Update the personal best clustering solution $P_i$.
   (b) end for
   (c) Find the global best solution $P_g$.
   (d) Update the cluster centers according to the velocity updating formula
       proposed in equation (3).
4. end for
```

2.4 Formal Concept Analysis

Formal concept analysis [9] is a mathematical formalism allowing to derive a concept lattice from a formal context constituted of a set of objects O, a set of attributes A, and a binary relation defined as the Cartesian product $O \times A$. The context is described as a table, the rows correspond to objects and the columns to attributes or properties and a cross in a table cell means that "an object possesses a property". Formal Concept Analysis (FCA) can be used for a number of purposes among which knowledge formalization and acquisition, ontology design, and data mining.

The *concept lattice* is composed of formal concepts, or simply concepts, organized into a hierarchy by a partial ordering (a subsumption relation allowing to compare concepts). Intuitively, a concept is a pair *(X, Y)* where $X \subseteq O$, $Y \subseteq A$, and X is the maximal set of objects sharing the whole set of attributes in Y and vice-versa. The set X is called the *extent* and the set Y the *intent* of the concept *(X, Y)*. The subsumption (or subconcept - superconcept) relation between concepts is defined as follows:

$$(X_1, Y_1) \prec (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 \text{ (or } Y_2 \subseteq Y_1). \tag{5}$$

Relying on this subsumption relation $\prec$, the set of all concepts extracted from a context is organized within a complete lattice, that means that for any set of concepts there is a smallest superconcept and a largest subconcept, called the *concept lattice*.

2.5 Integrative Clustering Approach Combining PSO and FCA

We propose herein an integrative clustering method that combines PSO and FCA in order to cluster datasets generated in multiple-experiment settings. It consists of two distinctive steps: *PSO-based clustering* and *FCA-based analysis*. Initially, the available microarray experiments are divided into groups of related (similar) datasets with respect to a predefined criterion. Subsequently, the hybrid (k-means and PSO) clustering algorithm as described in Section 2.3 [5] is applied to each group of experiments. This produces a list of different clustering solutions, one for each group. Next these solutions are pooled together and further analyzed by employing FCA which generates a single clustering solution for the whole data compendium of multiple experiments. FCA produces a concept lattice where each concept represents a subset of genes that belong to a number of clusters. The concepts compose the final disjoint clustering partition.

A detailed explanation of the distinctive phases of the proposed algorithm combining PSO and FCA is given below.

Initialization Phase. Assume that a particular biological phenomenon is monitored in several high-throughput experiments under a few different conditions. In this way, a list of different data matrices will be produced, one per experiment. Suppose that N different genes are in total monitored by all the different experimental datasets.

Initially, the available gene expression matrices are divided into r groups of related (similar) datasets with respect to some predefined criterion, *e.g.* the used synchronized method, or the expression similarity between the matrices. Then the number of cluster centers is identified for each group separately. As discussed in [11,22], this can be performed by running the selected clustering algorithm on each dataset for a range of different numbers of clusters. Subsequently, the quality of the obtained clustering solutions needs to be assessed in some way in order to identify the clustering scheme which best fits the datasets in question. For example, some of the internal validation measures that are presented in Section 3.2

(Silhouette Index or Connectivity) can be used as validity indices to identify the best clustering scheme. Suppose that k_i cluster centers are determined for each group i $(i = 1, 2, \ldots, r)$.

PSO-Grouped Clustering. The hybrid clustering algorithm explained in Section 2.3 is applied to each group of related experiments i $(i = 1, 2, \ldots, r)$ separately. The latter will generate a list of r different clustering solutions, one per each group, *i.e.* a set of k_i different clusters will be produced for each group i $(i = 1, 2, \ldots, r)$. Suppose that K $(K = k_1 + \ldots + k_r)$ different clusters in total are produced by all the different groups.

FCA Analysis. As discussed above, the N studied genes are grouped by the PSO-grouped clustering algorithm into K clusters. As mentioned in Section 2.4, FCA is a principled way of automatically deriving a hierarchical conceptual structure from a collection of objects and their properties. The approach takes as input a matrix (referred as the formal context) specifying a set of objects and the properties thereof, called attributes. In our case, a (formal) **context** consists of the set G of the N studied genes, the set of clusters $C = C_1, C_2, ..., C_K$ produced by the clustering step, and an indication of which genes belong to which clusters. Thus the context is described as a matrix, with the genes corresponding to the rows and the clusters corresponding to the columns of the matrix, and a value 1 in cell *(i, j)* whenever gene i belongs to cluster C_j. Subsequently, a (formal) **concept** for this context is defined to be a pair *(X, Y)* such that

- $X \subseteq G$ & $Y \subseteq C$ & every gene in X belongs to every cluster in Y
- for every gene in G that is not in X, there is a cluster in Y that does not contain that gene
- for every cluster in C that is not in Y, there is a gene in X that does not belong to that cluster.

The family of these concepts obeys the mathematical axioms defining a **concept lattice**. The constructed lattice consists of concepts where each one represents a subset of genes, all belonging to the same subset of clusters. The set of all concepts partitions the genes into a set of disjoint clusters.

On extremely large datasets the proposed integrative clustering method is expected to be computationally intensive. However, the computational cost can be drastically reduced by first performing some advanced filtering or features selection in order to remove noisy data and preserve lower number of potentially relevant genes for clustering.

3 Experimental Setup

3.1 Microarray Datasets

The proposed clustering algorithm has been validated on benchmark datasets where the true clustering is known. These datasets have been composed by gene expression time series data obtained from a study examining the global cell-cycle control of gene expression in fission yeast *Schizosaccharomyces pombe* [20]. The study includes eight independent time-course experiments synchronized respectively by:

1. elutriation: three independent biological repeats (*elu1*, *elu2*, *elu3*);
2. cdc25 block-release: two independent biological repeats, of which one in two dye-swapped technical replicates (*cdc25-1*, *cdc25-2.1*, *cdc25-2.2*) and in addition, one experiment in a sep1 mutant background (*cdc25-sep1*);
3. a combination of both methods: elutriation and cdc25 block-release (*elu-cdc10*) as well as elutriation and cdc10 block-release (*elu-cdc25*).

Thus, nine different expression test sets are available. In the preprocessing phase the rows with more than 25% missing entries have been filtered out from each expression matrix and any other missing expression entries have been imputed by the DTWimpute algorithm [23]. In this way nine complete matrices have been obtained.

Rustici *et al.* identified 407 genes as cell-cycle regulated [20]. These have been subjected to clustering which resulted in the formation of 4 separate clusters. Subsequently, the time expression profiles of these genes have been extracted from the complete data matrices and thus nine new matrices have been constructed. Note that some of these 407 genes were removed from the original matrices during the preprocessing phase, *i.e.* each dataset may have a different set of genes. Next, the nine datasets have been divided into three groups with respect to the used synchronization method. The overlapping genes within each group are as follows: a subset of 286 common genes in the elutriation datasets, a subset of 350 common genes in the cdc25 block-release datasets and a subset of 364 common genes in the datasets synchronized by the combination of both methods. Subsequently, the genes that are not presented in the intersection of the datasets of each group have been removed. As a result of this nine new matrices which form our benchmark datasets have been constructed. Notice that the nine different dataset contain 374 different genes in total.

The test datasets have been normalized by applying a data transformation method aiming at multi-purpose data standardization and inspired by gene-centric clustering approaches as proposed in [4].

3.2 Cluster Validation Measures

One of the most important issues in cluster analysis is the validation of the clustering results. Essentially, the cluster validation techniques are designed to find the partitioning that best fits the underlying data, and should therefore be regarded as a key tool in the interpretation of the clustering results. Since none of the clustering algorithms performs uniformly best under all scenarios, it is not reliable to use a single cluster validation measure, but instead to use at least two that reflect different aspects of a partitioning. In this sense, we have implemented two different validation measures for estimating the quality of the clusters:

1. Connectivity: for assessing connectedness;
2. Silhouette Index (SI): for assessing compactness and separation properties of a partitioning.

Connectivity. Connectivity captures the degree to which genes are connected within a cluster by keeping track of whether the neighboring genes are put into the same cluster [12]. Let us define $m_{i(j)}$ as the jth nearest neighbor of gene i, and let $\chi_{im_{i(j)}}$ be zero if i and j are in the same cluster and $1/j$ otherwise. Then for a particular clustering solution $C_1, C_2, \ldots, C_k$ of matrix M, which contains the expression values of m genes (rows) in n different experimental conditions or time points (columns), the connectivity is defined as

$$Conn(c) = \sum_{i=1}^{m} \sum_{j=1}^{n} \chi_{im_{i(j)}}.$$

The connectivity has a value between *zero* and *infinity* and should be *minimized*.

Silhouette Index. Silhouette index reflects the compactness and separation of clusters [19]. Suppose $C_1, C_2, \ldots, C_k$ is a clustering solution (partition) of matrix M, which contains the expression profiles of m genes. Then the SI is defined as

$$s(k) = \frac{1}{m} \sum_{i=1}^{m} (b_i - a_i) / \max\{a_i, b_i\},$$

where a_i represents the average distance of gene i to the other genes of the cluster to which the gene is assigned, and b_i represents the minimum of the average distances of gene i to genes of the other clusters.

The values of Silhouette Index vary from *-1* to *1* and *higher value* indicates better clustering results.

4 Validation Results and Discussion

In this section, the performance of the proposed integrative clustering method on the benchmark datasets is presented. The standard k-means, the hybrid (combination of k-means and PSO) clustering approach from Section 2.3 and the proposed integrative (combination of PSO and FCA) clustering algorithms are executed in order to generate clustering solutions on each of the considered nine microarray matrices. The quality of these solutions is evaluated using two cluster validation measures: Silhouette Index (SI) and Connectivity. These cluster validation measures have been implemented in C++. The PSO-based clustering algorithm has been implemented in Java. The publicly available open source machine learning software WEKA[3] is used by this implementation for the particle initialization and for the gene assignment to the different clusters.

[3] http://www.cs.waikato.ac.nz/ml/weka/

Initialization Phase. Initially, the nine test datasets are divided into three groups with respect to the used synchronized method:

1. elutriation datasets: *elu1*, *elu2*, *elu3*;
2. cdc25 block-release datasets: *cdc25-1*, *cdc25-2.1*, *cdc25-2.2*, *cdc25-sep1*;
3. datasets synchronized by the combination of both methods: *elu-cdc10*, *elu-cdc25*.

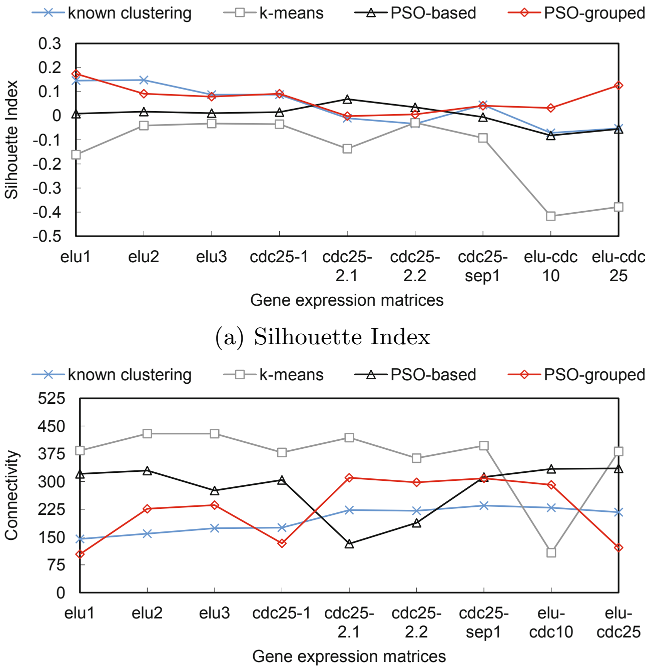
Then the number of cluster centers is identified for each group. As discussed in [11], [22], this can be performed by running the selected clustering algorithm on each dataset for a range of different numbers of clusters. Thus the k-means clustering algorithm is executed for values of k between 2 and 10 on each dataset. Subsequently, the quality of the obtained clustering solutions is assessed by using the Connectivity and SI as validity indices. We search for the values of k for which a significant local change in value of the index occurs [11]. The selected optimal number of clusters for the three groups of experiments is as follows: elutriation datasets: $k = 4$; cdc25 block-release datasets: $k = 6$, and the combined ones: $k = 5$.

PSO-Based Clustering. Next the PSO-based hybrid clustering algorithm (see Section 2.3) is executed on each group of experiments separately. It is run for 500 iterations with $w = 0.72$ and $c_1 = c_2 = 1.49$. These values have been chosen to ensure good convergence [17]. Notice that 15 different clusters (elutriation: clusters 0-3, cdc25 block-release: clusters 4-9 and combination of both: clusters 10-14) in total are produced by the three groups.

Figure 1 compares the SI and Connectivity values produced by the standard k-means, the known clustering solution published in [20], the hybrid PSO-based clustering algorithm considered in Section 2.3 and the PSO-grouped version of the latter algorithm described in Section 2.5 on the individual matrices. Note that the SI and Connectivity values for the PSO-based algorithm are obtained by using the global best solution found among the 9 different experimental matrices, those for the PSO-grouped algorithm are generated by using the global best solutions found separately for each group of datasets (elutriation, cdc25 block-release and combined), while the values for the k-means are produced by using the clustering solutions generated for each of the corresponding individual datasets.

As it can be seen in Figure 1, the SI values produced by the PSO-based and PSO-grouped algorithms outperform for all the nine experiments the values obtained for the k-means algorithm. Similar superior performance of the PSO-based and PSO-grouped algorithms in comparison to the k-means can be observed for the Connectivity scores with the single exception of *elu-cdc10.* The k-means result produced on the latter experiment appears to be an outlier of the Connectivity scores obtained for the rest of the experiments. There is no any obvious explanation of this phenomenon. It may be due to the experiment specific characteristics.

According to the SI indices, the PSO-grouped algorithm clearly outperforms the known clustering solution. However, the Connectivity index provides less



(a) Silhouette Index

(b) Connectivity Index

Fig. 1. Comparison of the SI (a) and Connectivity (b) values generated by the known clustering solution published in [20], and those obtained by applying the standard k-means, the PSO-based hybrid algorithm and the PSO-grouped version of the latter algorithm on the 9 different experiments. The SI and Connectivity values for the PSO-based algorithm are obtained by using the global best solution found among the 9 different experimental matrices, those for the PSO-grouped algorithm are generated by using the global best solutions found separately for each group of datasets (elutriation, cdc25 block-release and combined), while the values for the k-means are produced by using the clustering solutions generated for each of the corresponding individual datasets.

conclusive results. In general, the PSO-grouped clustering solution is better than the known one in 70% of the experiments under the SI validation index and respectively, in 35% of the test datasets under the Connectivity index. The PSO-grouped version also exhibits better performance than the PSO-based clustering algorithm in 80% of the experiments under both validation measures.

FCA Analysis. The gene partitions produced by the clustering step are further analyzed by applying FCA using publicly available tool [4]. We have created a context that consists of the set of 374 studied genes and the set of 15 clusters produced by the clustering step. It is described as a binary matrix, with the genes corresponding to the rows and the clusters corresponding to the columns. Subsequently, a lattice of 109 concepts for this context is generated (see Figure 2). Thus the FCA step partitions the benchmark gene set in 83 disjoint clusters (concepts) in total since the rest of the concepts appear to be empty. However, a number of 27 concepts are singleton sets and only the following seven concepts have cardinality above 10: {1, 6, 10}, {0, 5, 13}, {1, 6, 12}, {2, 4, 10}, {0, 5, 10}, {1, 8, 12}, {2, 6, 10}. It is interesting to notice that all the concepts connecting three clusters (46 such concepts exist) are not empty sets and in addition, they all contain clusters produced by each of the three groups of experiments.

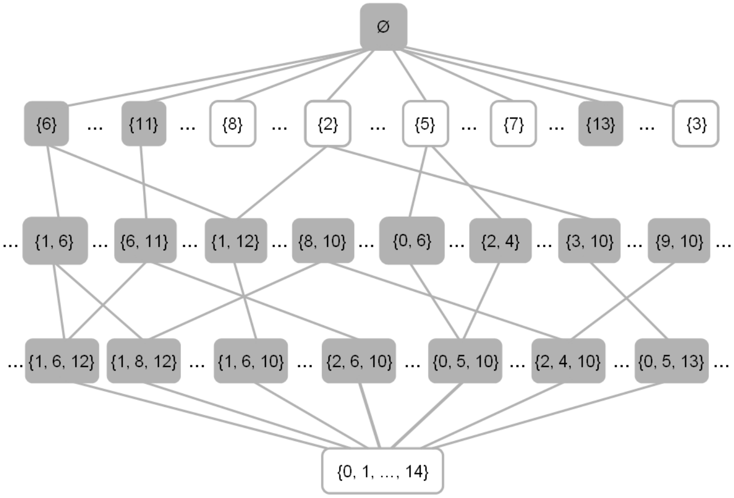


Fig. 2. Part of the generated concept lattice visualizing the seven concepts discussed above and their subconcepts and superconcepts

Each of the above listed seven concepts was subjected to analysis with the BiNGO tool [16], in order to determine which Gene Ontology categories are statistically overrepresented in each concept. The results are generated for a cutoff

[4] http://www.iro.umontreal.ca/~galicia/features.html

p-value of 0.05 and Benjamini and Hochberg (False Discovery Rate) multiple testing correction. For each gene concept a table is generated consisting of five columns: (1) the GO category identification (GO-id); (2) the multiple testing corrected p-value (p-value); (3) the total number of genes annotated to that GO term divided by total number of genes in the test set (cluster frequency); (4) the number of selected genes versus the total GO number (total frequency); and (5) a detailed description of the selected GO categories (description).

Only 5 of the seven FCA concepts (see above) have been assigned GO categories by the BiNGO tool: {0, 5, 13}, {1, 6, 12}, {2, 4, 10}, {1, 8, 12}, and {2, 6, 10}. Concretely:

- concept {0, 5, 13} contains 23 genes annotated to 4 GO categories (all have cluster frequency 78.2%), which point out to (cellular) response to stress and stimulas;
- concept {1, 6, 12} contains 18 genes connected with 5 GO categories (only 3 have total frequency $>$ 0.0%), all reffering to the regulation of sister chromatid cohesion and segregation;
- concept {2, 4, 10} contains 14 genes associated with about 100 GO categories (25% of these have total frequency = 0.0%), majority of which refer to regulation of different biological processes including cell-cycle;
- concept {1, 8, 12} contains 12 genes annotated to 22 GO categories (16 have total frequency $>$ 0.0%) dominated by RNA metabolic processing related categories;
- concept {2, 6, 10} contains 11 genes connected with 19 GO categories (10 have total frequency $>$ 0.0%), most of which refer to cell-cycle control or regulation of DNA replication.

Table 1. Cluster implications. The top row presents the number of genes contained in the premise clusters (second row). The clusters of the conclusion are listed in the third row and their corresponding number of genes in the forth row. The percentage of premise genes contained in the conclusion is given in the last row.

# Genes	24	14	15	6	7	4	1	1	1	1	1	1	1
Premise	0 13	14	3	2 13	3 14	2 14	0 7	2 5	2 8	2 11	3 6	3 7	9,11
	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Conclusion	5	4	4	6	4	4	11	10	10	6	12	12	1
# Genes	23	12	12	6	6	4	1	1	1	1	1	1	1
% Genes	96	86	80	100	86	100	100	100	100	100	100	100	100

Beside the generated concepts and the lattice diagram one can examine the implications between attributes (in our case the different gene clusters) valid in a context. Table 1 presents the specific dependencies extracted by the ConExp tool[5] between the clusters of the three groups of experiments. The *premise* defines a gene lattice and the *conclusion* specifies the dependent lattice that holds

[5] http://conexp.sourceforge.net/

for a high percentage of the genes in the premise. This implication describes that if a certain gene is present in the *premise* clusters, it is also found (with some exceptions) in the cluster from the *conclusion*. For instance, 23 out of the 24 genes present in clusters 0 (elutriation dataset) and 13 (combination of elutriation and cdc25 block-release dataset) are present in cluster 5 (cdc25 block-release dataset). Using these implications the genes occurring in the same clusters can be replaced by one representative gene. In addition, the genes that can be obtained as a result of the intersection of some other genes can be removed by ConExp reducing the 374 studied genes to a selection of 50. For example, if we consider the fourth column of Table 1, all the six genes belonging to clusters 2, 6, 13 are replaced by one representative gene. Further gene number 2, belonging to clusters 1, 10, can be obtained as a result of the intersection of genes 3 and 20 respectively, belonging to clusters 1, 6, 10 and 1, 7, 10 and therefore, we can remove gene 2. The so described reduction operation does not change the structure of the constructed lattice as the reduced concept lattice is isomorphic to the original one.

5 Conclusion

We have proposed an integrative clustering method which combines Particle Swarm Optimization and Formal Concept Analysis for deriving a clustering solution for multiple gene expression matrices. The performance of the proposed clustering algorithm has been evaluated on a test set of 9 time series expression datasets obtained from a study examining the global cell-cycle control of gene expression in fission yeast *Schizosaccharomyces pombe*. The presented in this article experimental results demonstrate that the proposed clustering algorithm is a robust data integration technique, which is able to produce good quality clustering solution that is representative for the whole test set. In addition, the employment of the FCA allows to perform a subsequent data analysis, which provides useful insights about the biological role of genes contained in the same FCA concepts. Our future work will focus on further exhaustive analysis of the composition and relationships between the different FCA concepts.

References

1. Alizadeh, A., et al.: Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403, 503–511 (2000)
2. Besson, J., Robardet, C., Boulicaut, J.-F.: Constraint-Based Mining of Formal Concepts in Transactional Data. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 615–624. Springer, Heidelberg (2004)
3. Besson, J., Robardet, C., Boulicaut, J.-F., Rome, S.: Constraint-based concept mining and its application to microarray data analysis. Intell. Data Anal. 9(1), 59–82 (2005)
4. Boeva, V., Tsiporkova, E.: A Multi-purpose Time Series Data Standardization Method. In: Sgurev, V., et al. (eds.) Intelligent Systems: From Theory to Practice. SCI, vol. 299, pp. 445–460. Springer, Heidelberg (2010)

5. Boeva, V., Hristoskova, A., Tsiporkova, E.: Clustering of Multiple DNA Microarrays through Combination of Particle Swarm Intelligence and K-means. In: Proceedings of the 6th International Conference on Computational Intelligence and Bioinformatics, Pittsburgh, USA, pp. 32–38 (2011)
6. Brazma, A., Gilks, W.R., Tom, B.D.M.: Fusing microarray experiments with multivariate regression. Bioinformatics 21(2), ii137–ii143 (2005)
7. Choi, J.K., et al.: Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 19, i84–i90 (2003)
8. Choi, V., Huang, Y., Lam, V., Potter, D., Laubenbacher, R., Duca, K.: Using formal concept analysis for microarray data comparison. Journal of Bioinformatics and Computational Biology 6(1), 65–75 (2008)
9. B. Ganter, G. Stumme, and R. Wille. Formal Concept Analysis: Foundations and Applications. *Lecture Notes in AI*, no. 3626, 2005, Springer-Verlag.
10. Golub, T., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
11. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17(2) (2001)
12. Handl, J., et al.: Computational cluster validation in post-genomic data analysis. Bioinformatics 21, 3201–3212 (2005)
13. Kaytoue-Uberall, M., Duplessis, S., Napoli, A.: Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes. In: Le Thi, H.A., Bouvry, P., Pham Dinh, T. (eds.) MCO 2008. CCIS, vol. 14, pp. 439–449. Springer, Heidelberg (2008)
14. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
15. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceeding of Fifth Berkeley Symp. Math. Stat. Prob., vol. 1, pp. 281–297 (1967)
16. Maere, S., Heymans, K., Kuiper, M.: BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. Bioinformatics 21, 3448–3449 (2005)
17. Omran, M., Engelbrecht, A., Salman, A.: Particle swarm optimization method for image clustering. Pattern Recognition and Artificial Intelligence (2005)
18. Potter, D.P.: A combinatorial approach to scientific exploration of gene expression data: An integrative method using Formal Concept Analysis for the comparative analysis of microarray data. Thesis dissertation, Department of Mathematics, Virginia Tech. (2005)
19. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational Applied Mathematics 20, 53–65 (1987)
20. Rustici, G., et al.: Periodic gene expression program of the fission yeast cell cycle. Nat. Genetics 36, 809–817 (2004)
21. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: Proceedings of IEEE Int. Conf. on Evolutionary Computation, pp. 69–73 (1998)
22. Theodoridis, S., Koutroubas, K.: Pattern recognition. Academic Press (1999)
23. Tsiporkova, E., Boeva, V.: Two-pass imputation algorithm for missing value estimation in gene expression time series. Journal of Bioinformatics and Computational Biology 5(5), 1005–1022 (2007)
24. Zhou, et al.: Functional annotation and network reconstruction through cross-platform integration of microarray data. Nature Biotechnology 23(2), 238–243 (2005)

Link Prediction Approaches for Disease Networks

Francesco Folino and Clara Pizzuti

National Research Council of Italy (CNR)
Institute for High Performance Computing and Networking (ICAR)
Via Pietro Bucci, 41C
87036 Rende (CS), Italy
{ffolino,pizzuti}@icar.cnr.it

Abstract. In the last years link prediction in complex networks has attracted an ever increasing attention from the scientific community. In this paper we apply link prediction models to a very challenging scenario: predicting the onset of future diseases on the base of the current health status of patients. To this purpose, a comorbidity network where nodes are the diseases and edges represent the contemporarily presence of two illnesses in a patient, is built. Similarity metrics that measure the proximity of two nodes by considering only the network topology are applied, and a ranked list of scores is computed. The higher the link score, the more likely the relationship between the two diseases will emerge. Experimental results show that the proposed technique can reveal morbidities a patient could develop in the future.

1 Introduction

Medical care research in the field of preventative medicine constitutes a fundamental activity to determine the risk of individuals to develop diseases, and to undertake the correct actions at the earliest signs of illness. Hospitals and physicians, every year, collect thousands of patient clinical histories that can be exploited to build prediction models by considering the *comorbidity relations* of patients, instead of results of laboratory tests. A comorbidity relationship between two illnesses exists whenever they appear simultaneously in a patient more than chance alone [9].

Recently, Davis et al. [4, 3] proposed to assign a patient with the list of diseases he has been affected during his life, and to represent him by a vector of diagnosed disease codes, defined by the *International Classification of Diseases, Ninth Revision, Clinical Modification* ICD-9-CM. They realized a collaborative assessment and recommendation engine that relies on collaborative filtering methodologies [17]. Prediction on patient's diseases are done by comparing the individual medical history with the medical histories of a set of similar patients. The similarity between two patients is defined by taking into account the random expectation of each diseases and the inverse frequency of each disease. Association analysis, clustering, and Markov chains have also been successfully applied and combined

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 99–108, 2012.

in [5–7] for disease prediction by exploiting the frequent sets of illnesses that contemporarily appear in patients. Steinhaeuser and Chawla [18] built a disease network where nodes are the diseases and edges connects those pairs of diseases appearing in the same patient. Structural properties of the network, such as degree distribution, hubs, diameter are then studied. Furthermore, they used a hybrid technique based on collaborative filtering and nearest neighbor classification to predict diseases.

In this paper, analogously to [9] and [18], we construct a comorbidity network and exploit the connections between diseases to build a prediction model. Differently from the described approaches, we propose the utilization of *link prediction methods* to generate a disease predictive model.

Link prediction is the problem of predicting the existence of a connection between two objects by considering the informations on the objects and the interactions already present in the network [8]. The problem aims at inferring those interactions among the members of a network that are likely to occur in the future. The disease prediction problem can thus be naturally formulated as a link prediction problem if objects are diseases and a predicted link (x, y) between two diseases x, y means that a patient affected by the illness x (y resp.) is likely to be effected also by disease y (x resp.) in the future. The search for these missing links is performed by applying *similarity-based algorithms* (see Section 3) that assign a score to the pair (x, y), that measures the proximity between x and y. Experimental results on a disease network built on a set of true diagnosis show that the approach can reveal morbidities a patient could develop in the future. It is worth to note that, to the best of our knowledge, this is the first attempt that exploits link prediction techniques to foresee emerging disease interactions in health-related networks.

The paper is organized as follows. The next section describes the data set of patients and the disease network representing the comorbidity relations present in this data set. Section 3 introduces the link predictions methods employed to infer a connection between two illnesses. Section 4 describes the evaluation method employed. Section 5 presents the results and evaluates them by comparing the precision of the approaches with respect to a random predictor. Finally, Section 6 concludes the paper.

2 Disease Network

The dataset T we deal with consists of 2541 patient medical records coming from two small towns of the south of Italy. Each patient is associated with the list of ICD-9-CM (*International Classification of Diseases, Ninth Revision, Clinical Modification*) codes of the illnesses he has been diagnosed during a period of about ten years (from 2000 to 2009). Only the first three digits (on the overall five) of ICD-9-CM codes expressing the general diagnosis have been considered. Even if some details can be missed, these three digits are sufficiently informative to study the disease correlations. Table 1 reports some details about the dataset at hand. In particular, the top-20 most frequent diseases per patient are highlighted.

Table 1. Top-20 most recurrent diseases per patient

Rank	**Disease**	% **of occurrences**
1	401 *(Hypertension)*	$33,16\%$
2	530 *(Diseases of esophagus)*	$22,02\%$
3	715 *(Osteoarthrosis and allied disorders)*	$21,54\%$
4	722 *(Intervertebral disc disorders)*	$17,72\%$
5	462 *(Pharyngitis)*	$14,81\%$
6	250 *(Diabetes mellitus)*	$14,26\%$
7	466 *(Acute bronchitis and bronchiolitis)*	$11,70\%$
8	733 *(Other disorders of bone and cartilage)*	$10,32\%$
9	724 *(Other and unspecified disorders of back)*	$8,90\%$
10	464 *(Acute laryngitis and tracheitis)*	$8,86\%$
11	721 *(Spondylosis and allied disorders)*	$8,07\%$
12	240 *(Simple and unspecified goiter)*	$8,07\%$
13	272 *(Disorders of lipoid metabolism)*	$6,58\%$
14	595 *(Cystitis)*	$6,54\%$
15	535 *(Gastritis and duodenitis)*	$6,42\%$
16	427 *(Cardiac dysrhythmias)*	$6,18\%$
17	600 *(Hyperplasia of prostate)*	$6,07\%$
18	300 *(Neurotic disorders)*	$5,04\%$
19	491 *(Chronic bronchitis)*	$4,84\%$
20	726 *(Peripheral enthesopathies and allied syndromes)*	$4,81\%$

The patient medical records contain important enlightenment regarding the co-occurrences of diseases affecting the same individual. A *comorbidity relationship* between two illnesses exists whenever they appear simultaneously in a patient more than chance alone [9]. In order to make discernible the correlations among the diseases contained in our dataset, a *disease network* whose nodes are the diseases and a link between two nodes occurs every time a comorbidity relation appears, i. e. when the couple of diseases affects at least one patient, is built. The edges are labelled with the number of patients showing both the illnesses.

The number of nodes is 492 and the number edges is 21676. Since many of the weights between two diseases is 1, the same statistical approaches proposed by Hidalgo et al. [9] to measure the strength of comorbidity relationships, and thus to discard those edges deemed less meaningful, have been adopted. The measures employed to quantify the strength between two sicknesses are the *Relative Risk* (*RR*) and the *ϕ-correlation*. The RR of observing a pair of diseases i and j appearing in the same patient is given by

$$RR_{ij} = \frac{CC_{ij}(N - CC_{ij})}{P_i P_j} \tag{1}$$

where CC_{ij} is the number of patients affected by both diseases, N is the total number of patients in the data sets, and P_i, P_j are the numbers of patients affected by diseases i and j, respectively. On the other hand, the *ϕ-correlation* is defined as

$$\phi_{ij} = \frac{CC_{ij}(N - CC_{ij}) - P_i P_j}{\sqrt{(P_i P_j (N - P_i)(N - P_j))}} \tag{2}$$

As pointed out in [9], the Relative Risk overestimates relations involving rare diseases and underestimates relationships between very common sicknesses. Instead, *ϕ-correlation* underestimates comorbidity between rare and frequent diseases, and accurately discriminates associations between illnesses of similar appearances. Thus, we built a network by selecting only the statistically significant edges having $RR > 20$, shortly named G_{RR}, and another one obtained by discarding all the edges having $\phi \leq 0.06$ named G_{ϕ}. Cumulatively, the resulting number of edges for G_{RR} was 2330, whereas for G_{ϕ} we obtained 7242 edges.

In the following we first describe the link prediction techniques, then we apply them to both G_{RR} and G_{ϕ}.

3 Link Prediction Metrics Based on Node Neighborhoods

Many methods have been proposed for link prediction [13]. Some of them are the so-called *similarity-based algorithms*, which assign a score $score(x, y)$ to a pair of nodes (x, y) of a generic network G by computing a proximity measure between nodes. The similarity indices differ according to the information used to define them. Often node features are not known, thus similarity is computed by considering only the network topology. A ranked list in decreasing order of $score(x, y)$ is thus generated, and the higher the score between to nodes, the higher the probability that the relationship between them will emerge in the future. In the following we consider local similarity indices based on the neighborhood of nodes [12], in particular we explore four measures: (1) *Common Neighbors*; (2) *Jaccard*'s coefficient; (3) *Adamic-Adar* index and (4) *Resource Allocation* index. A brief description of these indices is given in the following.

Let x be a node of a network G and $\Gamma(x)$ denote the set of neighbors of x in G. Many of the most used approaches leverage on the idea that two nodes x and y will have more chance to form a link in the future if their sets of neighbors $\Gamma(x)$ and $\Gamma(y)$, respectively, overlap. In [11, 2] this principle has been exploited for modeling the growth of networks, i.e., a link (x, y) is more likely to form if (x, z) and (z, y) already exist for some z in the network G.

Common-Neighbors. The most immediate implementation of the idea that two nodes are more likely to be connected if they share more common neighbors, consists in defining the score as

$$score_{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{3}$$

In [14] this quantity has been applied in the context of collaboration networks. In particular, it has been proven that a correlation between the number of common neighbors of x and y at time t and the probability that they will collaborate in the future exists.

Jaccard's Coefficient. Jaccard's coefficient [10] is a very common used similarity metric in information retrieval field [16]. It computes the probability that two nodes x and y share a feature f, where f is a randomly selected feature of either x or y. If the neighbors of a node are considered as its features, the Jaccard's coefficient can be analytically expressed in this way:

$$score_J(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{4}$$

Adamic-Adar Index. In order to decide how correlated two personal home pages are, Adamic and Adar [1] defined a similarity measure on the base of their shared features. This index refines the simple counting of common neighbors by assigning to rarer features a more heavy weight. Thus,

$$score_{AA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(k(z))} \tag{5}$$

where $k(z)$ is the degree of node z.

Resource-Allocation Index. In [19], Zhou et al. introduced a new proximity metric by taking inspiration from a real physical process: the resource allocation in networks [15]. Let x and y indicate a pair of nodes not directly connected. The idea is that x sends some resource to y while their common neighbors act as transmitters. In the simplest case, x has a unit of resource and can send part of it to y by averagely distributing the amount to all its neighbors. The greater the amount of resources y receives from x, the more likely a link between x and y will form. Therefore, the score between x and y can be defined as

$$score_{RA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} \tag{6}$$

In the next section we will apply the similarity indices described to the two disease networks G_{RR} and G_ϕ, and will test their effectiveness in the health context.

4 Experimental Setup

For testing the accuracy of the predictors over G_{RR} and G_ϕ, we adapted the approach proposed in [12] to the medical context. We first randomly divided the dataset T in two parts T_{train} and T_{test}, containing the 90% and the 10% of patients, respectively. Let $G(V, E)$ be the disease network (i.e., either G_{RR} or G_ϕ) representing the dataset T. Then, two sub-networks $G_{train}(V_{train}, E_{train})$ over T_{train}, and $G_{test}(V_{test}, E_{test})$ over T_{test} are built. While E_{train} represents known information and it is used to generate a prediction model, the probe set $E_{new} = E_{test} - E_{train}$ consists of those couples of diseases that are present in the test set but not in the train set, that the predictor should discover because they have high probability to occur in the future. Clearly, $E = E_{train} \cup E_{new}$ and $E_{train} \cap E_{new} = \emptyset$.

For the sake of clearness, Table 2 reports the number of links in E_{train} and E_{new} for both G_{RR} and G_ϕ networks. We recurred to a precision measure to assess the accuracy of the predictors, which is defined as follows.

Table 2. Number of links in the disease networks

Disease Network	$\|\mathbf{E_{train}}\|$	$\|\mathbf{E_{new}}\|$
G_{RR}	2096	234
G_{ϕ}	6517	725

Let V be the set of vertices of the disease network built over the whole patient dataset T. For each pair of nodes, $x, y \in (V \times V) - E_{train}$, every predictor described in Section 3 assigns a score $score(x, y)$. As explained in the previous section, this value measures the similarity between nodes x and y, and the greater the score, the higher the probability that the corresponding link, missing at current time, will emerge in the next future. Then, all the predicted links are sorted in decreasing order according to their scores into a list L_p. Moreover, let n denote the cardinality of E_{new} and L_p^n the top-n high-scored links in L_p. Therefore, the precision can be defined as:

$$Precision = \frac{|L_p^n \cap E_{new}|}{|E_{new}|} \tag{7}$$

In order to better understand how the precision metric works, let us consider the following example.

Example. Figure 1(a) shows a simple network G having five nodes, seven existent edges and three nonexistent ones (i.e., (1,2), (1,4) and (3,4)). For learning the prediction model, we need to select some edges from the whole graph G to form the training set $E_{train} = \{(1,5),(2,3),(2,4),(2,5),(3,5)\}$ (represented in the network G_{train} in Figure 1(b)). The remaining links define the probe set $E_{new} = \{(1,3),(4,5)\}$ (as in the graph G_{new} of Figure 1(c)).

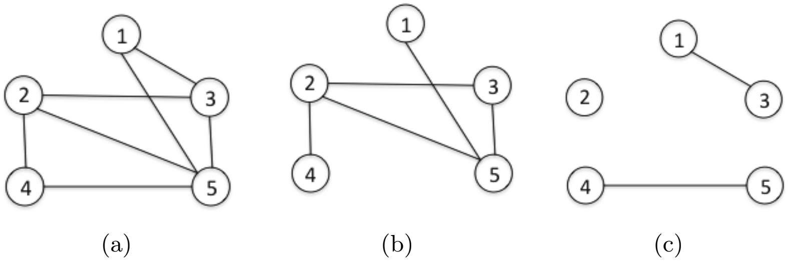


Fig. 1. The three networks involved in the Precision calculus: (a) G, (b) G_{train}, and (c) G_{new}

Among those presented in Section 3, let *Jaccard* be the predictor used to compute the score $score_J(x, y)$ for each couple of nodes in the set of all non-observed links $N_l = (V \times V) - E_{train} = \{(1,2),(1,3),(1,4),(3,4),(4,5)\}$. It is worth to note that the prediction algorithm can only exploit information contained in the training graph G_{train}. The sets Γ of neighbors for each node are the following: $\Gamma(1) = \{5\}, \Gamma(2) = \{3,4,5\}, \Gamma(3) = \{2,5\}, \Gamma(4) = \{2\}, \Gamma(5) = \{1,2,3\}$. Then, by computing the $score_J$ (see Eq. (4)) for each link in N_l, it is easy to verify the results are those reported in Table 3.

Table 3. Jaccard's score for each non-observed link

Edge (x,y)	$\mathbf{Score_J(x,y)}$
$(1,2)$	0.33
$(1,3)$	0.5
$(1,4)$	0
$(3,4)$	0.5
$(4,5)$	0.33

Since $n = 2$ is the cardinality of the probe set E_{new}, thus the list of predicted links L_p^2 will contain the two highest-scored edges $\{(1,3),(3,4)\}$. Clearly, if we look at the Figure 1(c), the intersection between L_p^2 and E_{new} is the link $(1,3)$). Therefore, by applying the formula expressed in Eq. (7), we finally obtain that the *Precision* value is equal to 0.5 □

5 Results

This section is devoted to evaluate the link prediction approach to infer links between diseases. We shall present the results and assess them on the base of the introduced metrics.

Table 4 reports the capability of each predictor to foresee the appearance of new (i.e., not known at prediction time) disease correlations in both G_{RR} and G_ϕ networks. Notice that the best prediction performances (marked in bold) are achieved by *Resource-Allocation* and *Adamic-Adar* predictors over the G_ϕ network. A motivation for this result can be found in the inherent properties of the metrics (i.e., RR and ϕ) used to construct the links in the networks. As a matter of fact, the network G_ϕ contains highly prevalent diseases with many connections across different ICD-9-CM categories [9]. Likely, this topological structure increases the probability of two diseases to share similar links in the network, and, accordingly, to raise up their similarity.

In order to evaluate the precision values obtained, we compared them against the results obtained by a baseline predictor which randomly selects pairs of diseases from those not correlated in the training set. Results in Table 5 show that the random predictor achieves a precision of 0.4274% over G_{RR} and 0.4138% over G_ϕ, respectively. In the same table, each predictor performance in terms of the factor of improvement w.r.t. the baseline is also reported. It is easy to note that while the results of different techniques are quite comparable, each of these significantly outperforms the random predictor by a factor of around $40 - 50$. In particular, both the *Adamic-Adar* and *Resource-Allocation* approaches obtain a notable 48.67 of improvement factor w.r.t. the baseline result over G_ϕ. This

Table 4. Precision for different predictors (in bold higher values)

Network	**Precision**			
	Common-Neighbors	*Jaccard*	*Adamic-Adar*	*Resource-Allocation*
G_{RR}	0.175213675	0.162393162	0.183760684	**0.188034188**
G_{ϕ}	0.191724138	0.180689655	**0.20137931**	**0.20137931**

Table 5. Improvement factor over random prediction

Predictor	**Disease Network**	
	G_{RR}	G_{ϕ}
Random	0.4274%	0.4138%
CommonNeighbors	41	46.33
Jaccard	38	43.67
Adamic-Adar	43	48.67
Resource-Allocation	44	48.67

behavior clearly suggests that there is some hidden information in the network topology that can be profitably exploited for prediction purposes.

Finally, we want to emphasize the practical usefulness of link prediction in inferring disease correlations unknown in the training set, but likely emerging in the patient's near future. To this purpose, Table 6 reports, as an example, few of the highest scored links predicted by the *Resource-Allocation* approach over the network G_{ϕ}. It is worth to note that the table contains disease correlations either totally new or known in the probe set only. In more details, the predictions ranked $\{1, 7, 8\}$ are new links, whereas the remaining ones were present in E_{test}. While the latter are implicitly validated, we miss an adequate medical knowledge to assess the meaningfulness of the former correlations. However, some correlations among general classes of diseases, recognized in the medical literature, are reported in Table 7. They can be used as a sort of background knowledge to validate the new links. The interesting aspect to observe is that, besides the correlations in Table 6 ranked $\{2, 3, 4, 5, 6, 9, 10\}$ (that are present in the test set), also the new links could be explained by the high-level interactions appearing in Table 7. For instance, the fourth row of Table 7 can explain the

Table 6. Top-10 links of L_p predicted by *Resource-Allocation* over G_{ϕ}

Rank	**Predicted Links**		**Score**
	Disease	Disease	
1	*Ischemic heart disease* (414)	*Iron deficiency anemias* (280)	0.886971
2	*Essential hypertension* (401)	*Disorders of parathyroid gland* (252)	0.858515
3	*Essential hypertension* (401)	*Disorders of iron metabolism* (275)	0.760756
4	*Old myocardial infarction* (412)	*Simple and unspecified goiter* (240)	0.760661
5	*Occlusion of cerebral arteries* (434)	*Simple and unspecified goiter* (240)	0.760578
6	*Acute myocardial infarction* (410)	*Thyroiditis* (245)	0.741699
7	*Duodenal ulcer* (532)	*Parkinson's disease* (332)	0.720294
8	*Osteoarthrosis and allied disorders* (715)	*Dermatomycosis* (111)	0.720055
9	*Other disorders of kidney and ureter* (593)	*Candidiasis* (112)	0.685907
10	*Phlebitis and thrombophlebitis* (451)	*Disorders of fluid/electrolyte...* (276)	0.680304

new discovered link of the 8th row of Table 6, between "Osteoarthrosis and allied disorder" and "Dermatomycosis".

Table 7. Most common disease correlations known in the medical literature

Known disease interactions	
`ICD-9-CM group`	`ICD-9-CM group`
Infectious and parasitic d. (001–139)	*Respiratory system d.* (460–519)
Infectious and parasitic d. (001–139)	*Genitourinary system d.* (580–629)
Infectious and parasitic d. (001–139)	*Skin and subcutaneous tissue d.* (680–709)
Infectious and parasitic d. (001–139)	*Musculosk. sys./conn. tissue d.* (710–739)
Endocr./nutr./metab./imm. d. (240–279)	*Circulatory system d.* (390–454)
Nervous system d. (320–359)	*Digestive system d.* (520–579)

This result confirms both the concrete relevance of link analysis when applied to the disease prediction context, and a more general behavior that is the main lesson we learned: a large number of new correlations are hinted by the topology of a disease network that actually contains plentiful latent information from which inferring future disease interactions.

6 Conclusions

The problem of predicting the onset of new disease correlations is not an easy task. In fact, there can be multiple causes that originate an illness, and the mechanisms underlying them are not clear. However, preventive medicine can improve the life quality since a patient can modify his habits and lifestyle, in order to prevent the appearance of probable, correlated future diseases. The paper presented an approach based on link predictions methods to identify diseases a patient could develop by considering the structural similarity between nodes of a disease network. Experiments on a real network showed that the network structure can reveal latent information useful for inferring new disease correlations. Future work will aim at applying more complex link prediction models that take into account also nodes characteristics and edges weights, information that is not considered in the similarity metrics exploited for our experiments.

Acknowledgements. This work has been partially supported by the project *MERIT : MEdical Research in Italy*, funded by MIUR.

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social Networks 25, 211–230 (2001)
2. Davidsen, J., Ebel, H., Bornholdt, S.: Emergence of a small world from local interactions: Modeling acquaintance networks. Phys. Rev. Lett. 88, 128701 (2002)

3. Davis, D.A., Chawla, N.V., Christakis, N.A., Barabási, A.L.: Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery 20, 388–415 (2010)
4. Davis, D.A., et al.: Predicting individual disease risk based on medical history. In: Proc. of the ACM Int. Conf. on Information and Knowledge Management, CIKM 2008, pp. 769–778 (2008)
5. Folino, F., Pizzuti, C.: A comorbidity-based recommendation engine for disease prediction. In: Proc of 23rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2010, pp. 6–12 (2010)
6. Folino, F., Pizzuti, C.: Combining Markov Models and Association Analysis for Disease Prediction. In: Böhm, C., Khuri, S., Lhotská, L., Pisanti, N. (eds.) ITBAM 2011. LNCS, vol. 6865, pp. 39–52. Springer, Heidelberg (2011)
7. Folino, F., Pizzuti, C., Ventura, M.: A Comorbidity Network Approach to Predict Disease Risk. In: Khuri, S., Lhotská, L., Pisanti, N. (eds.) ITBAM 2010. LNCS, vol. 6266, pp. 102–109. Springer, Heidelberg (2010)
8. Getoor, L., Diehl, C.P.: Link mining: a survey. SIGKDD Explorations 7, 3–12 (2005)
9. Hidalgo, C.A., Blumm, N., Barabási, A.L., Christakis, N.A.: A dynamic network approach for the study of human phenotypes. PLoS Computational Biology 5(4) (2009)
10. Jaccard, P.: Bulletin de la Societe Vaudoise des Science Naturelles 37 (1901)
11. Jin, E.M., Girvan, M., Newman, M.E.J.: Structure of growing social networks. Phys. Rev. E 64, 046132 (2001)
12. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the 12th International Conference on Information and Knowledge Management, CIKM 2003, pp. 556–559 (2003)
13. Lü, L., Zhou, T.: Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390(6), 1150–1170 (2009)
14. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Phys. Rev. E 64, 025102 (2001)
15. Ou, Q., Jin, Y.-D., Zhou, T., Wang, B.-H., Yin, B.-Q.: Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. Phys. Rev. E 75, 021102 (2007)
16. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc. (1986)
17. Shardanand, U., Maes, P.: Social information filtering: algorithms for automating word of mouth. In: Proc. of ACM Conf. on Human Factors in Computing Systems, CHI 1995, pp. 210–217 (1995)
18. Steinhaeuser, K., Chawla, N.V.: A network-based approach to understanding and predicting diseases. In: Social Computing and Behavioral Modeling (2009)
19. Zhou, T., Lü, L., Zhang, Y.-C.: Predicting missing links via local information. The European Physical Journal B - Condensed Matter and Complex Systems 71(4), 623–630 (2009)

Toward a Semantic Framework for the Querying, Mining and Visualization of Cancer Microenvironment Data

Michelangelo Ceci[4], Fabio Fumarola[4], Pietro Hiram Guzzi[3],
Federica Mandreoli[6], Riccardo Martoglia[6], Elio Masciari[1],
Massimo Mecella[2], and Wilma Penzo[5]

[1] ICAR-CNR, Italy
[2] La Sapienza University, Italy
[3] Magna Graecia University, Italy
[4] University of Bari
[5] DEIS - University of Bologna, Italy
[6] DII - University of Modena and Reggio Emilia
ceci@di.uniba.it, fabiofumarola@gmail.com, hguzzi@unicz.it,
{federica.mandreoli,martoglia.riccardo}@unimo.it, masciari@icar.cnr.it,
mecella@dis.uniroma1.it, wilma.penzo@unibo.it

Abstract. Over the last decade, the advances in the high-throughput omic technologies have given the possibility to profile tumor cells at different levels, fostering the discovery of new biological data and the proliferation of a large number of bio-technological databases. In this paper we describe a framework for enabling the interoperability among different biological data sources and for ultimately supporting expert users in the complex process of extraction, navigation and visualization of the precious knowledge hidden in such a huge quantity of data. The system will be used in a pilot study on the Multiple Myeloma (MM).

1 Introduction

The emergence of affordable high-performance computers, and the high-throughput omic technologies are the basis of several projects aiming at building new public molecular profile databases and data repositories on clinical cancer and cultured cancer cell lines. Using such new public databases, biomedical researchers can i) publish their data and results making them available to the scientific community, and ii) use the in-lab produced and public data to study a drug candidate, a gene or a disease state in a biological system in order to verify hypothesis and generate new knowledge. Major examples of public "bio-technological databases and repositories" are: the National Center for Biotechnology Information (NCBI) located in United States, the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). They host data about genome (sequences, maps, chromosomes, assemblies, and annotations), proteins, nucleotides, genes (reference sequences, maps, pathways,

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 109–123, 2012.

variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources), relationships between phenotype and genotype and more than 21 million citations from biomedical literature. Subsequently, other interesting project initiatives have come out, each of which with the goal of providing useful information with respect to a particular viewpoint of complex biological systems. Gene Ontology (GO) project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data. GO allows users to query and to extract knowledge from the built ontology. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database stores a collection of online databases dealing with genomes and enzymatic pathways. The KEGG pathway data bank records networks of molecular interactions in the cells, and variants of them specific to particular organisms. The DrugBank, the KEGG DRUG and the Chemical Entities of Biological Interest (ChEBI) databases offer different kinds of bioinformatics and cheminformatics resources that combine detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The U.S. National Cancer Institute (NCI) with the NCI-60 database offers tools for storing, querying and, downloading molecular profile data of 60 diverse human cancer cell lines used since 1990 to screen compounds for anticancer activity. In addition to the above described ones, a particularly interesting project is the Connectivity Map (CMap), which is born with the challenge of establishing relationships among diseases, physiological processes, and the action of drugs (small molecules). The CMap provides a solution to this problem by:

- describing all biological states (physiological, disease, or induced with a chemical or genetic construct) in terms of genomic signatures;
- creating a large public database of signatures of drugs and genes;
- developing pattern-matching tools to detect similarities among these signatures.

Prevalently, in the literature, CMap is queried by researchers using signatures obtained from comparative gene expression analysis (e.g. disease compared with normal state, treated drugs versus not treated ones) to identify drug response profile that either correlate or anti-correlate with it. When a signature is derived from clinical samples representing a disease, the discovered connections represent a list of sample-drugs which either mimics or reverts the disease signature, while a signature obtained from drug-treated cells can be used to retrieve a list of chemical compounds with similar effect. CMap plays a central role because it relates, from a genomic point of view, diseases, genes' functions and drugs' actions according to the same language.

The added value of the framework we propose in this paper is obtained by linking out CMap with the various types of data and partial knowledge stored in different data banks, including those cited above. By performing a comprehensive analysis of databases, data repositories, and ontologies, our aim is not to replicate existing data, but to design and develop a Web delivery system which:

1. enables the interoperability among the queryable data sources;
2. captures the different kinds of relationships that exist among them;
3. reinforces the cooperation of heterogeneous and distributed data bank sources for the query processing target;
4. supports the users in the complex process of extraction, navigation and visualization of the knowledge hidden in such a huge quantity of data.

In particular, to facilitate interoperability (1), we will focus on the normalization problem by creating a semantic layer linking the data sources (2). On top, innovative algorithms and techniques for querying (3), mining and visualizing data, models and statistics will enable the extraction of new knowledge (4). This would support bio-medical researchers in analyzing tumors microenvironments in order to understand them and identify relationships among tumors, the effect of drugs and the patients' biodiversity. Such relationships are of particular interest for drug repositioning and for the identification of novel compounds able to overcome resistance or revert it. The system will be used in a pilot study on the Multiple Myeloma (MM), an incurable malignant plasma cell disease with an incidence of 5 per 100,000 inhabitants, and for that in NCBI GEO are submitted around 6658 samples. MM locates primarily to the bone marrow (BM) in multiple niches that provide a microenvironment which promotes tumor survival. In particular all the functionalities of our project will be exploited to understand the relationships of MM with other tumors, to understand mechanisms of drug resistance in MM Cells (MMCs) to 4 drugs (Dexamethasone, Bortezomib, High Dose Melphalan, Lenalidomide) in current use, to elucidate the contribution of the tumor environment in conferring drug resistance in MMCs, to identify novel compounds able to overcome resistance or revert it in MMCs to drugs in current use, to identify a set of candidate gene products for extensive study of their role in drug resistance in MMCs.

2 Background

In this section, we will discuss the preliminary concepts mandatory to our system implementation that emerged after a deep analysis performed with the support of biological data experts. In particular for each topic we will discuss in detail the state of the art and the limitations of current proposals that lead us to the system architecture described in Section 3.

2.1 The Connectivity Map

Biological data features made traditional approaches to data analysis inadequate for their efficient analysis and understanding thus a great effort has been devoted to research on these fields. A main problem is to integrate data coming from different data sources and obtained with different acquisition technologies. In this perspective, the Connectivity Map (CMap) plays an important role[34]. Briefly, it is a freely available search engine that may be used to retrieve information about diseases, drugs and gene expression levels. It stores raw data about the gene expression published in different papers, and information about the impact of different drugs on the expression level. The main utility of CMap, as evidenced in recent papers, is the possibility to generate in silico hypothesis, by analyzing gene expression data. The usage of the CMap as a starting point for our activities is motivated by the increasing interest of the scientific community in the prediction of novel associations among diseases, physiological processes, and the action of small therapeutic molecules [3]. Several papers in the literature proposed different solutions in order to extend the information stored in the Connectivity Map with other bio-technological databases. Examples are [25],[35],[18],[39] and [44] where the CMap is extended by the definition of new genomic signatures for diseases and chemical compounds extracted from GEO Dataset and Series, new distance measures for chemical compounds and pathologies based on protein network interaction and PUBMED abstracts. However, all the above contributions represent ad-hoc extensions of the CMap. At the moment, there is no systematic approach which allows us to extend and integrate the information stored in the CMap with data available in other bio-tecnological database such as NCBI (Gene, Geo, Pubmed), ArrayExpress, Gene Ontology, KEGG, and Drug Bank. Finally, when normalizing data a crucial activity is data de-duplication. A proposed approach consists in the adoption of a hierarchical clustering method [29], equipped with a suitable record matching scheme, that leverages accurate field-wise similarity metrics to match corresponding record tokens. To overcome efficiency and effectiveness problems, the adoption of a hash-based index has been proposed[6].

2.2 Interoperability

The dataspace principles have been recently introduced in the literature[20] as a data management abstraction alternative to data integration where, unlike fully integrating heterogeneous data sources, the coexistence of data, which is autonomously modeled and loosely connected through relationships for sharing purposes, is supported. This new paradigm better fits the data scenario envisioned by the project, where a full control of data sources is not always available. Currently, a huge amount of biological data can be naturally represented by graphs. Several works have been proposed on the graph query problem, which has been extensively studied in the context of a graph database consisting of a set of relatively small graphs, while little attention has been paid to the context of a single large graph[47], which is the context common to most biological networks. The major challenge in this scenario is to reduce the number of pairwise subgraph isomorphism checkings, since subgraph isomorphism is known to be a NP-complete problem. A number of graph indexing techniques, where different structural patterns are examined to help prune the candidate search space, have been proposed to address this challenge (e.g. [24]). As to the problem of supporting approximate graph matching, which is a decisive feature to support queries on heterogeneous data, only few works have been proposed[48]. Much work has been done w.r.t. the problem of querying both databases and mining datasets. As to the relational data model, a data mining language based on the principles of closure (the results of a query can be further manipulated) and cross-over between data and rules[27] has been introduced. Extensions to the object-oriented data model[12] have also been reported in the literature. However, approximate querying where graph-based data and mining datasets co-exist is a research issue that has never been addressed. Also, given the different syntactic and semantic representations and the massive scale of the datasets, checking whether multiple data instances are actually the same entity is a very challenging problem. The proposed record linkage techniques are either id-based, when ids are available, or apply syntactic approximations on the data. Most of these approaches are integration-oriented and are not suitable to the data coexistence scenario envisioned by the project. Very few works present on-the-fly techniques[28] that will be considered as the starting point to propose the hybrid approach needed to dynamically support the different connection alternatives for data sources.

2.3 Data Mining

As regards the mining activities on biological data a key task is the discovery of groups of genes whose gene expressions are simultaneously altered by one or more pathologies. This analysis would provide useful information for drug repositioning[2], that is, understanding whether drugs typically used for treating some specific tumors can be used for treating other tumors since they report at a normal state the same genes (e.g. it is a recent finding the fact that chemical compounds typically used in treatments for tumors have positive effects on patients suffering from the Alzheimer's disease[7]). A possible improvement is the exploitation of co-clustering discovery approaches, which are the most suited tools to identify clusters of objects of different nature (in this case genes-patologies). Indeed, the application of co-clustering techniques in the biological context is not new[22]. However, most of the existing approaches focus on the algorithms that, if applied to large datasets, present the problem of a high number of extracted co-clusters. Moreover, most of the existing approaches suffer from the impossibility of extracting overlapping co-clusters (in our case a gene can be involved in several regulation networks)[8]. Finally, existing co-clustering algorithms do not consider possible relationships that involve first class objects considered in the analysis (i.e. genes and pathologies) or relationships between these first class objects and other objects (possibly of a different type, such as functional pathways and mRNA). In order to overcome these limitations recent studies in Collective Classification[42] has been presented. They allow to take into account possible autocorrelation in the data. Another interesting research line related to our project is the study of evolution of pathologies through short time series analysis techniques. Unlike traditional time series analysis, whose main problem is related to the length of time series, short time series are characterized by very few temporal points. This is a characteristic of our extended CMAP. According to [14], relevant tasks for short time series are: classification, clustering and anomaly detection. The use of interactive visual interfaces for cycles identification to perform classification of sequences, to perform comparisons among sequences as well as to perform pattern matching and temporal pattern search have been addressed in literature [43], but very few of them can be used for short time series, a proposal on this field is in [46]. Recent studies focus on the definition of tools for the identification of the pathology stage on the basis of expression gene values. To this end, approaches of collective classification will be particularly studied[45] due to their peculiarity to handle the autocorrelation aspects typically present in data organized in network/graph form. Indeed, it has been proved in the literature that the co-occurrence of autocorrelation with high-density neighborhood of data could bias the selection of features in the task of relational classification. Evolutionary algorithms (EA)[15] are heuristics that mimic the processes of natural evolution in order to solve global search problems. They differ from more traditional optimization techniques in that they involve a search starting with a "population" of solutions (i.e. a string of bits), not with a single point. Recombination, crossover and mutation operators are used to generate new solutions that are biased towards different regions of the search space. Genetic programming (GP)[32] is an extension of genetic algorithms (GAs) that iteratively evolves a population of (typically) trees of variable size, by applying variation operators. The use of hybrid techniques, i.e. EAs and data mining ensembles, together with efficient implementations and with new models of distributed computations enables these kinds of algorithm to cope with hard classification problems. Bagging and boosting, introduced in [41] and [16] are well known ensemble techniques that repeatedly run a learning algorithm on different distributions over the training data.

2.4 Semantic Tagging and Ontologies

Semantic relationships among biological entities are actually a growing research field. These relationships are usually derived from biological knowledge encoded into biological ontologies. There exist many active projects that aim to organize such knowledge into structured vocabularies of concepts and taxonomies of concepts themselves. For instance Gene Ontology (geneontology.org/)[23] stores concepts about the localization, processes and function of genes and gene products, while Protein Ontology (proteinontology.org.au/)[37] contains concepts related to proteins and Disease Ontology (diseaseontology.sourceforge.net) relationships among diseases and genes. In computational biology, ontologies are often used to annotate biological concepts, i.e. to associate to a biological concept such as gene BRCA1 its description using only concept from a biological field[11]. The use of ontologies to support modeling and querying of biological data has been explored in [19]. In recent years, many approaches for the evaluation of the similarity of two or more concepts belonging to the same ontology have been developed. Thus, starting from two entities that are annotated with terms belonging to the same ontology, it is possible to define a semantic similarity by the similarity of the concepts used for annotating them. In this way, it is possible to define all pairwise similarities of entities belonging to the same domain and annotated with the same ontology (even if this approach may be extended in the case of different ontologies) [19]. The whole set of entities and their similarities may be efficiently represented into a single comprehensive model by using graph theory [17]. Finally, there exist also different computational approaches developed both for the knowledge extraction of biological networks (e.g. clusterings) [1],[5] and for concepts belonging to the same ontology (e.g. semantic similarity measures)[19].

2.5 Visual Query Languages

A Visual Query System (VQS) is a system that uses a visual representation for both the domain of interest and the related requests on it. Two forms of query creation (SQL and of Query By Example (QBE)) have been recently compared through experiments [26]. The authors found that the time requested for query formulation applying a QBE-based approach was shorter than the time requested using a SQL approach. Interestingly, there were no remarkable differences regarding the accurateness of the queries between the two approaches. An example of diagram-based Visual Query Language is QBD*[4]. A framework that allows both intensional and extensional queries, developed following the paradigm Query By Browsing (QBB), is described in [40]. QBI[36] is a pure iconic Visual Query Language, which provides tools for an intensional browsing of databases. Finally, as stated by [31], Visual Analytics is "the science of analytical reasoning facilitated by interactive visual interfaces", which means to help the decision making process by turning the information overload into an opportunity. Human perception plays an important role in such a tool because a visualization which does not consider cognitive principles may lead to misunderstandings and wrong interpretations of the data.

3 Our Approach

As previously discussed, a fundamental challenge that arises throughout biomedicine is the need to establish relations among diseases, physiological processes, and the action of small-molecule therapeutics. Our goal is to introduce an end to end system (whose architecture is depicted in Fig. 1) that would provide a technological support to this issue by exploiting the CMap and additional data sources. In this way, the bio-medical researchers will be able to study Cancer microenvironments in order to understand their specificities and the effect of drugs considering the patients' biodiversity.

As stated before, such relationships can give better insights about tumors as well as can be used for drug repositioning and for the identification of novel compounds able to overcome resistance or revert it in to drugs in current use. Our aim is to offer the following functionalities within a user friendly Web delivery system:

- Identification of the data repositories and databases which relate to the CMap;
- Normalization and interoperability of the identified databases and the CMap;
- Extraction of useful knowledge from the data by means of data Mining techniques;

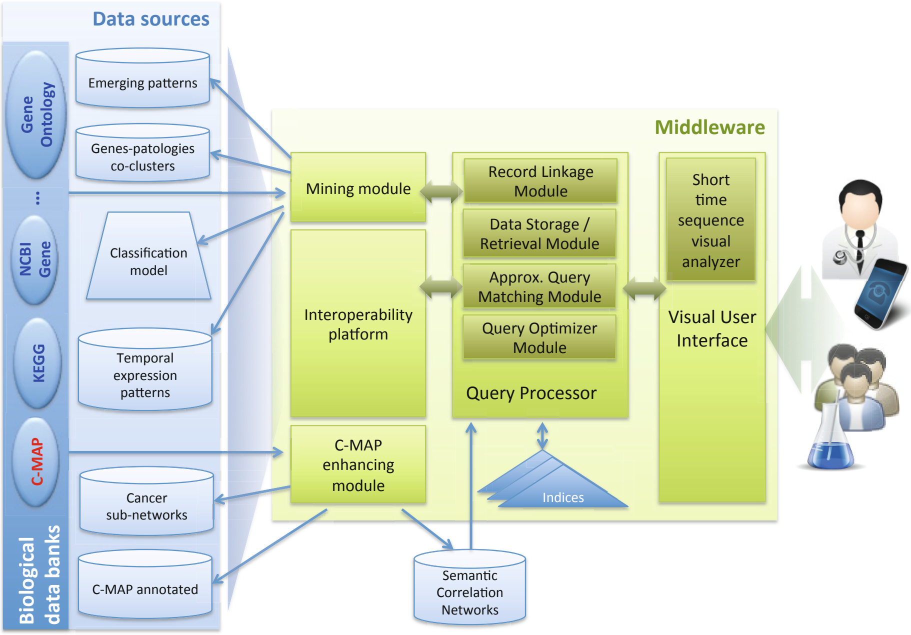


Fig. 1. The system architecture

- Semantic tagging of CMap;
- Querying of the extended CMap, the identified data repositories and the extracted knowledge as a unique dataspace;
- Querying CMap and extracted knowledge by means of a Visual Query Language.

To this end, we aim to give innovative answers to the problems that are mandatory for these objectives, by applying several methodologies and techniques.

3.1 Bio-technological Data Gathering

Our first goal is to identify the data repositories which can be combined with the Connectivity Map, keeping in mind that the final goal is to allow bio-medical researchers to navigate the stored knowledge as well as to formulate new hypotheses based on the information stored in the bio-technological data repositories and databases. In this perspective, the central role of the Connectivity Map in this project is motivated by the increasing interest of the scientific community in the prediction of novel association among diseases, physiological processes, and the action of small therapeutic molecules. As a matter of fact, as mentioned above several works in the literature aim at extending the information stored in the CMap. However, they represent ad-hoc extensions. Our goal, instead, is to provide a systematic approach to extend the CMap and make the information it stores interoperable with data available in other bio-tecnological database such as NCBI (Gene, Geo, Pubmed), ArrayExpress, Gene Ontology, KEGG, and Drug Bank as reported in Fig. 1. The interoperability issues arising from accessing such a wide set of data sources are discussed in Section 3.3.

3.2 Data Mining and Semantic Information Extraction

By taking into account the requirements coming from bio-medical researchers we exploit several mining algorithms in order to extract knowledge on Cancer microenvironments from the selected datasources. The goal is to better understand tumors and to identify relationships among tumors, the effect of drugs and the patients' biodiversity. As stated before, such relationships can be used for drug repositioning and for the identification of novel compounds able to overcome resistance or revert it in drugs in current use. Due to both the novelty and the nature of the these problems, ad hoc data mining algorithms are used. Indeed, a significant benefit is the identification of semantic relatedness among domain-specific entities. In particular, the results of the data mining algorithms are used to enrich/confirm ontologies that can be used to support semantic-based querying. On the other hand, semantically tagged data can be used to identify relationships that can be considered in the mining processes. Accordingly, we introduce the notion of semantic relatedness that relates CMap entities. By exploiting co-clustering approaches, we discover groups of genes whose gene expressions are simultaneously altered by one or more pathologies [9]. To this purpose, hierarchical and non-hierarchical co-clustering techniques are exploited. Moreover, in order to characterize and describe pathologies (or classes of pathologies) on the basis of the variability of genomic signatures observed in gene products, network based emerging patterns discovery algorithms [10] are used. Furthermore, we analyze evolution of pathologies through short time series analysis techniques [13]. To this end, both visual data mining and temporal patterns extraction algorithms are defined. Finally, in order to identify the pathology stage on the basis of expression gene values, collective classification algorithms and ensemble-based algorithms are exploited. While in the case of collective classification [42] it is possible to handle the autocorrelation (according to which "closer" objects are more related than "furthermost" ones), typically present in data organized in network/graph form, in ensemble-based classification [38] different learning models will be combined together (ensemble) in order to define the final model. All the above mentioned mining algorithms will be implemented in the "Mining module" of our system (see Fig. 1). As regards data correlation analysis, starting from data stored in the CMap, the selected databases and the ontologies, a Semantic Correlation Network will be built. This semantic graph will be used to extract sub-networks related to Cancer through the application of network analysis algorithms such as network alignment algorithms [33], clustering [29], and pattern extraction algorithms [21]. The outcomes of this activity isimplemented in the "CMap enhancing module" of the Web delivery system.

3.3 Biotechnological Data Modeling and Management

Once all the data repositories have been identified, additional problems raise. Indeed, the biological data to be analyzed are heterogeneous both in their type and format, since they come from several data sources exhibiting different schema. Moreover, another kind of information that is particularly useful for our goal is the knowledge provided by the mining activities. Once again, it differs from the biological data not only for the format but mainly for the adopted model as it refers to a mining model rather than operational ones. On the other hand, all the above mentioned data sources are inherently connected, thus the availability of normalization and interoperability solutions that would allow analysis tools to deal with information coming from different sources in a unified way is crucial. In addition, solutions to enrich the CMap with the information gathered from the other biological data sources are necessary to use semantics to search or browse its data. Finally, a flexible query model is necessary that would allow stakeholders to easily query the knowledge in the data sources in a uniform way and to get useful results for analysis purposes. Therefore, the main challenges related to this goal are:

1. Extension of the Connectivity Map with semantic information encoded into ontologies;
2. Normalization and interoperability of the set of data sources;
3. Definition of techniques effectively and efficiently supporting the querying of the data sources.

As regards the first challenge, RDF annotations to the CMap entities with the support of the selected ontologies will be introduced. The output will be stored in a relational database (*C-Map annotated*, see Fig. 1) containing both entities and functional annotations extracted from ontologies whereas the methods will be implemented in an ad hoc module, called *Annotation Module*. It will create the first version of *C-Map annotated*, then it will periodically update it by searching for new annotations that can be extracted from publicly available databases.

The second challenge will be dealt with the aim of providing a technological platform to the full interoperability among the selected data banks and the outputs of the various tasks: C-Map annotated; the sub-networks related to Cancer; the genes-pathologies co-clusters, the disease (emerging) patterns, temporal expression patterns (extracted from short time sequences) and the disease classification model. The interoperability platform will then support the co-existence of all these sources, each of which can participate both as internal source or as external one, through two languages, a *Data Delivery Definition Language* (DDDL) for source specification and a *Mapping Language* (ML) for inter-source relationship specification. Specifically, the interoperability platform will include low-level repository functionalities, including: a) (for all data sources) storage and retrieval of data source associated descriptions, including DDDL and ML; b) (only for complete access sources) storage /retrieval / update of the data itself; c) (only for external sources) storage and retrieval of wrapping patterns.

The third challenge is faced through 1) the creation of a flexible query language (equipped with an approximate query matching model) that would allow stakeholders to easily query the system and to get useful results, 2) the definition of algorithms and data structures for approximate query answering that would ensure good performances under different system conditions. The language allows users to specify queries as graphs of biological concepts, biological entities (data instances), predicates on biological entities, and labeled relationships among them. Moreover it will extend the classical comparison operators with ad-hoc operators to query both data and mining models. Query samples that could be specified are "Find all genes that are up-regulated and whose localization is similar to Nucleolus and function is similar to receptor-binding", "Find all the groups of similar genes whose localization is different from Nucleolus that are down-regulated under the effect of drug X", and many others. Once a query is issued to the system, the query processor module will approximate the query on the dataspace by: 1) defining a query plan that selects the involved sources through the interoperability platform, 2) sending to each selected source the appropriate query, collecting and merging the query results through the application of record linkage techniques [30].

3.4 Definition of the Web Delivery System for the Easy-Access to the Bio-technological Data and the Derived Knowledge

All the project results will be made available to the research community through the Web delivery system as depicted in Fig. 1. The system will be implemented as a Service Oriented Infrastructure according to which Web-services enabling both access to data and usage of the defined algorithms will be provided. An important feature is the user-friendliness of the whole prototype for potential users. The Web delivery system will then be made accessible by means of a Visual User Interface module that provides biological data experts with a rich user experience during the usage of the tool, both in the querying phase and in the result manipulation phase. The main objective is then the definition of a visual query language specifically targeted to biological data sources analysis and of appropriate visualization techniques. In order to fully explain the goal we intend to obtain, consider the following example: the query and visualization interface of CMAP – http://www.broadinstitute.org/cmap/. By using this tool, a query basically consists in providing a signature file and searching for connected objects. The main difficulties for users are in the text-based syntax of the signature file, which almost requires a kind of programming capabilities, as the syntax should be rigorous, etc. In particular, nowadays the way of writing a query is to create an Excel file and to insert specific values into the columns, according to the given sheet format. Conversely, in our system a graphical interface is developed, in which the user, through drag & drop of the basic elements needed for building a signature (to be taken from a palette available to the user), is able to visually write such a signature and to use it for querying the system. In the same way, currently the results of the queries are viewed in a table format, and then for each of them a click allows for opening the related specification (again an Excel file). Conversely, a graph-based visualization is envisioned, in which results are shown as nodes of a graph, and the edges represent relationships (e.g., due to sharing of some objects in the structure, etc.). Different colors, thickness of the edges, etc. convey specific semantics. A more natural interaction modality will also allow for the use of the interface/tool by users equipped with modern devices, such as tablets, during their normal operations in laboratories, etc. and therefore do not impose the use of a desktop.

4 Case Study: Supporting Bio-Medical Researchers in the Study of the Multiple Myeloma through the Web Framework

The web framework will be used in a pilot study on the Multiple Myeloma (MM), an incurable malignant plasma cell disease. MM is particularly interesting since it allows researchers to concentrate on microenvironments which promote tumor survival. Main challenges related to this goal can be tackled in our system, such as:

1. Better understanding the Multiple Myeloma (MM) by expoiting semantic record linkage techniques. This is performed by identifying other blood tumors that are semantically related to the MM not only on the basis of information stored in the CMap, but also on the basis of information stored in other datasources such as PubMed (literature) and Gene Ontology;
2. Provide the bio-medical researchers with a tool for querying the dataspace identified for the MM pilot study;
3. Support the work of bio-medical researchers with Data Mining techniques.

To this purpose, the aim is to allow bio-medical researches to avoid wasting time and funds for the in-vitro verification of potentially meaningless hypothesis by their testing with in silico techniques. In other words, on the basis of results obtained through the application of Data Mining techniques, it is possible to drive the process of hypothesis generation in:

- understanding the correlation of the MM with other tumors in terms of gene expressions modifications;
- defining a characterization of the MM in terms of genes;
- analyzing the short time evolution of the pathology;
- generating classification models to automatically identify MM on the basis of gene expressions modifications and additional information stored in other datasources. This analysis might help the drug repositioning task and the identification of novel compounds able to overcome resistance or revert it in drugs in current use;
- provide the bio-medical researchers with semantic network analysis techniques. Indeed, since drug resistance in the MM is possibly related to drug resistance in other (blood) tumors, we use semantic network analysis techniques in order to identify the semantic distance between MM and each other (blood) tumor (again, to support drug repositioning).

5 Conclusion

In this paper we presented a system for biological data normalization and interoperability devoted to information extraction, data querying and knowledge dissemination for supporting biomedical specialists in the analysis of cancer microenvironments. The system is modular and is tailored on the biological data features in order to make it easy to use and provide useful information to the domain experts. The system will be used for assisting bio-medical researcher in the analysis of information related to Multiple Myeloma.

References

1. Aittokallio, T., Schwikowski, B.: Graph-based methods for analysing networks in cell biology. Brief Bioinform. 7(3), 243–255 (2006)
2. Ashburn, T.T., Thor, K.B.: Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery 3, 673–683 (2004)
3. Boutros, P.C.: Fun with microarrays part iii: Integration and the end of microarrays as we know them. Hypothesis 6(1) (2008)
4. Catarci, T., Santucci, G.: Query by diagram: A graphical environment for querying databases. In: Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, May 24-27, p. 515. ACM Press (1994)
5. Ciriello, G., Guerra, C.: A review on models and algorithms for motif discovery in protein interaction networks. Briefings in Functional Genomics and Proteomics 7(2), 147–156 (2008)
6. Costa, G., Manco, G., Ortale, R.: An incremental clustering scheme for data deduplication. Data Min. Knowl. Discov. 20(1), 152–187 (2010)
7. Cramer, P.E., Cirrito, J.R., Wesson, D.W., Lee, C.Y.D., Karlo, J.C., Zinn, A.E., Casali, B.T., Restivo, J.L., Goebel, W.D., James, M.J., Brunden, K.R., Wilson, D.A., Landreth, G.E.: Apoe-directed therapeutics rapidly clear ß-amyloid and reverse deficits in ad mouse models. Science 335(6075), 1503–1506 (2012)
8. Deodhar, M., Gupta, G., Ghosh, J., Cho, H., Dhillon, I.S.: A scalable framework for discovering coherent co-clusters in noisy data. In: Pohoreckyj Danyluk, A., Bottou, L., Littman, M.L. (eds.) ICML. ACM International Conference Proceeding Series, vol. 382, p. 31. ACM (2009)
9. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. In: ISMB, pp. 145–154 (2002)
10. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: KDD, pp. 43–52 (1999)
11. Plessis, L.D., Kunca, N., Dessimoz, C.: The what, where, how and why of gene ontology a primer for bioinformaticians. Briefings in Bioinformatics (2011)
12. Elfeky, M.G., Saad, A.A., Fouad, S.A.: ODMQL: Object Data Mining Query Language. In: Dittrich, K.R., Oliva, M., Rodriguez, M.E. (eds.) ECOOP-WS 2000. LNCS, vol. 1944, pp. 128–140. Springer, Heidelberg (2001)
13. Ernst, J., Bar-Joseph, Z.: Stem: a tool for the analysis of short time series gene expression data. BMC Bioinformatics (2006)
14. Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics 21(suppl. 1), i159–i168

15. Fogel, D.B.: Evolutionary computation - toward a new philosophy of machine intelligence, 3rd edn. Wiley-VCH (2006)
16. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: ICML, pp. 148–156 (1996)
17. Golumbic, M.C.: Algorithmic Graph Theory and Perfect Graphs. Academic Press, New York (1980)
18. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: Predict: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. 7 (2011)
19. Guzzi, P.H., Mina, M., Guerra, C., Cannataro, M.: Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in Bioinformatics (2011)
20. Halevy, A.Y., Franklin, M.J., Maier, D.: Principles of dataspace systems. In: PODS, pp. 1–9 (2006)
21. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2000)
22. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics 18(suppl. 1), S145–S154 (2002)
23. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., Tonellato, P., Jaiswal, P., Seigfried, T., White, R.: The gene ontology (go) database and informatics resource. Nucleic Acids Res. 32, 258–261 (2004)
24. He, H., Singh, A.K.: Closure-tree: An index structure for graph queries. In: ICDE, pp. 38–49 (2006)
25. Hu, G., Agarwal, P.: Human disease-drug network based on genomic expression profiles. PLoS One 4(8), e6536 (2009)
26. Hvoreckya, J., Drlikb, M., Munk, M.: The effect of visual query languages on the improvement of information retrieval skills. Procedia - Social and Behavioral Sciences 2(2), 717–723 (2010)
27. Imielinski, T., Virmani, A.: Msql: A query language for database mining. Data Min. Knowl. Discov. 3(4), 373–408 (1999)
28. Ioannou, E., Nejdl, W., Niederée, C., Velegrakis, Y.: On-the-fly entity-aware query processing in the presence of linkage. PVLDB 3(1), 429–438 (2010)
29. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31 (September 1999)
30. Karmel, R., Gibson, D.: Event-based record linkage in health and aged care services data: a methodological innovation. BMC Health Services Research (2007)
31. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual Analytics: Scope and Challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
32. Koza, J.R.: Genetic Programming On the Programming of Computers by Means of Natural Selection. MIT Press (1992)

33. Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., Przulj, N.: Topological network alignment uncovers biological function and phylogeny. J. of the Royal Society (2010)
34. Lamb, J.: The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer 7(1), 54–60 (2007)
35. Li, J., Zhu, X., Chen, J.Y.: Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. PLoS Comput. Biol. 5(7), e1000450 (2009)
36. Massari, A., Pavani, S., Saladini, L., Chrysanthis, P.K.: Qbi: Query by icons. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, p. 477. ACM Press (1995)
37. Natale, D., Arighi, C., Barker, W., Blake, J., Chang, T.-C., Hu, Z., Liu, H., Smith, B., Wu, C.: Framework for a protein ontology. BMC Bioinformatics 8 (2007)
38. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198 (1999)
39. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. Journal of Computational Biology a Journal of Computational Molecular Cell Biology 18(2), 133–145 (2011)
40. Polyviou, S., Evripidou, P., Samaras, G.: Query by browsing: A visual query language based on the relational model and the desktop user interface paradigm. In: The 3rd Hellenic Symposium on Data Management (2004)
41. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning 37(3), 297–336 (1999)
42. Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective classification in network data. AI Magazine 29(3), 93–106 (2008)
43. Shah, M., Corbeil, J.: A general framework for analyzing data from two short time-series microarray experiments. IEEE/ACM Trans. Comput. Biol. Bioinformatics 8(1), 14–26 (2011)
44. Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A., Sage, J., Butte, A.J.: Discovery and preclinical validation of drug indications using compendia of public gene expression data. Science Translational Medicine 3(96), 96–77 (2011)
45. Stojanova, D., Ceci, M., Appice, A., Džeroski, S.: Network Regression with Predictive Clustering Trees. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 333–348. Springer, Heidelberg (2011)
46. Wang, X., Wu, M., Li, Z., Chan, C.: Short time-series microarray analysis: Methods and challenges. BMC Systems Biology 2 (2008)
47. Zhao, P., Han, J.: On graph query optimization in large networks. Proc. VLDB Endow. 3(1-2), 340–351 (2010)
48. Zhu, L., Ng, W.K., Cheng, J.: Structure and attribute index for approximate graph matching in large graphs. Inf. Syst. 36(6), 958–972 (2011)

Argumentation to Represent and Reason over Biological Systems

Adam Wyner[1,*], Luke Riley[1], Robert Hoehndorf[2], and Samuel Croset[3]

[1] Department of Computer Science, University of Liverpool, Liverpool, UK
[2] Department of Genetics, University of Cambridge, Cambridge, UK
[3] European Bioinfomatics Institute, Cambridge, UK
adam@wyner.info

Abstract. In systems biology, networks represent components of biological systems and their interactions. It is a challenge to efficiently represent, integrate and analyse the wealth of information that is now being created in biology, where issues concerning consistency arise. As well, the information offers novel methods to explain and explore biological phenomena. To represent and reason with inconsistency as well as provide explanation, we represent a fragment of a biological system and its interactions in terms of a computational model of argument and argumentation schemes. Process pathways are represented in terms of an argumentation scheme, then abstracted into a computational model for evaluation, yielding sets of 'consistent' arguments that represent compatible biological processes. From the arguments, we can extract the corresponding processes. We show how the analysis supports explanation and systematic exploration in a biology network.

Keywords: argumentation, systems biology, computational methods.

1 Introduction

Systems biology is an inter-disciplinary field that emphasizes the analysis of whole biological systems and the interactions occurring within them. Instead of reducing the behavior of biological systems to that of its parts, biological phenomena are studied as components in a network of interrelated processes that span multiple domains and levels of granularity [14]. Computational methods in systems biology rely on the construction of models that can predict the behavior of biological systems, the integration of large amounts of data derived from multiple sources, experimental methods and domains as well as the study of networks of interactions between the components of biological systems [13]. Biological networks represent components of biological systems and their interactions [3], which have been crucial in the analysis of protein-protein interactions [12], side-effects [15] and human disease [11].

Several large curated knowledge repositories have been created to store information about these interactions [6,10,15], and the application of high-throughput

[*] Corresponding author.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 124–138, 2012.

technologies in molecular biology further contributes to the rapid increase of information about interactions that occur in biological systems. It is an ongoing challenge to efficiently represent, integrate and analyze the wealth of information that is now being created in biology. Integration and analysis of data in systems biology is challenging on different levels: first, for curators of scientific data and knowledge bases, it is difficult to identify inconsistencies from source textual materials; second, data repositories may be incomplete, error-prone, or inconsistent; third, combining multiple repositories is difficult and can lead to inconsistencies; and fourth, the wealth of information that is now available requires entirely new analysis methods that can identify explanations and supporting evidence for biological phenomena.

Novel computational representations and implementations that can automatically represent and reason over biological phenomena can facilitate curation of databases, data retrieval, integration of data across multiple domains and levels of granularity, and assemble alternative or competing interpretations for experiment results. Here, we explore the possibility to represent biological systems and their interactions in terms of *argumentation theory*, a computational model of argument which is used to represent and reason with *inconsistency*. We will use the example of biochemical pathways to illustrate how components and their interactions in biological systems can be represented using the framework of argumentation theory.

Abstract argumentation frameworks (AFs) are a means to represent and reason with inconsistencies. AFs use graphs of nodes and arcs, where the nodes represent abstract arguments, having no internal structure, and the arcs represent *attacks* between the arguments [7]. Over complex networks of arguments, we can calculate *extensions*, which are sets of arguments which are mutually compatible, though the intersection of the sets has incompatible arguments. Where arguments are related to component propositions, then extensions are semantic models of a domain. Adding or subtracting arguments (and their corresponding attacks) from an AF gives rise to alternative extensions. Such frameworks have been widely developed to handle non-monotonic reasoning.

However, abstract arguments are not useful for representing instantiated arguments, that is, arguments with some internal structure or content such as in logical syllogisms or in presumptive reasoning argumentation schemes [20]. Such instantiated arguments appear in knowledge bases, which themselves are widespread for many domains. Some efforts have been made to relate abstract to instantiated arguments [1,2,5,19,23]; these tend to have domain specific forms. One of the strengths of argumentation schemes is that they provide an 'explanation' or 'justification' of a conclusion; where schemes are chained together, a rich explanation is provided.

In this paper, we develop an argumentation scheme to support reasoning for biomolecular pathways; a reaction in a pathway is represented in terms of propositions in an argument that has premises, exceptions, a rule, and a conclusion, where the inhibitions or perturbations are represented as exceptions. Instantiated schemes or chains of schemes represent arguments. Such arguments can stand in attack relations, where the conclusion of one argument is the negation of part of another argument. Where we abstract from the schemes and give the attacks, we can express the network in an argumentation framework and calculate extensions. In this way, we can represent knowledge bases of biomolecular pathways, reason with *inconsistency* that may arise either between knowledge bases or as knowledge in a domain grows, and explain outcomes. More importantly, argumentation theory provides a novel approach to analyze chains of complex interactions in biological systems, evaluate the consequences of defects in these systems, and possibly provide a model of therapeutic strategies for complex diseases with a molecular basis. While an extensive evaluation of our framework based on real biological data is future work, we exercise the analysis with respect to a small, worked example derived from an existing pathway knowledge base, illustrating how biologically significant questions about interaction networks can be restated as operations in an argumentation framework.

The paper has the following sections. Our materials and queries for biomolecular pathways are indicated in section 2. In section 3, we outline argumentation frameworks and instantiated argumentation. The formal language in which our *biomolecular argumentation scheme* is expressed is given in section 4; the language underpins the scheme. The scheme is given in section 5, then instantiated. Additional instantiations are then shown to represent a fragment of a given biomolecular pathway. In effect, a knowledge base for biomolecular pathways is translated into a format that suits tableau reasoning. The advantage of the instantiated scheme is that it gives specific *locations* for attack. To exercise the analysis, we abstract from the particulars of the instantiated scheme to represent an AF and its extensions. Concerning the introduction or removal of particular elements of the AF, we calculate extensions with respect to alternative attacks, which may be interpreted as *in silico* experiments. We discuss related work in section 6 and future work in 7.

2 Systems Biology Background

The behavior of complex biological entities such as cells and organisms is the result of interacting entities across multiple scales of granularity. For example, the type of proteins that are expressed in a cell will determine the function of the cell through a complex network of interactions such as positive and negative regulations. Depending on which proteins are present in the cell (i.e., expressed and then translated), the functions of cells (and, on a higher level of granularity, tissues and organs) change. Not all proteins are expressed simultaneously; instead, modules of protein interaction networks are stable sets of proteins that usually are expressed together in order to result in stable functioning of a cell. Furthermore, these proteins do not interact randomly, but are based on stable pathways that evolved over time. Pathways are chains of interactions that have been identified as significant because they result in a particular biological function or a product that is crucial for the functioning of a cell. Several pathway databases aim to capture this information. A ‘normative’ state of a cell such as expressed by a pathway or a network of interacting proteins may be disrupted, either pathologically in the case of a disease or disorder, or by the introduction of a drug or another biological agent such as a microRNA (which negatively regulate the transcription of mRNAs). Depending on which of the proteins are present in a cell and how they interact, the physiology of the cell, and subsequently the tissue and organ of which the cell is a part, changes.

Important questions about interaction networks in systems biology include the identification of stable functional modules, i.e., entities and interactions that may occur simultaneously and in parallel without conflict [3]. Once we are able to identify such modules, we can investigate the effect of changing the normative behavior of such interaction networks. For example, we can investigate the effect of inhibiting particular interactions or interacting entities, such as when we introduce drugs or regulatory elements such as microRNAs that either selectively inhibit the activity of molecules or disrupt the occurrence of interactions between molecules [4]. In a more complex scenario, we may want to identify an entity (e.g., a drug) that, when added to an interaction network, can achieve a desired outcome while at the same time minimizing adverse reactions resulting from this introduction. Ultimately, these operations would be evaluated against experimental data such as gene expression experiments.

Our basic assumption is that argumentation frameworks can not only provide the means to reason with inconsistent knowledge bases, but also provide the means to analyze interactions in biological systems. In particular, we aim to test the hypothesis that some types of interaction networks in biology can be represented as networks of arguments, and that notions such as *consistency*, *rule*, *premise* and *attack* from argumentation theory correspond to constituents of biological systems and their underlying laws.

3 Argumentation Frameworks and Instantiated Argumentation

To represent and reason with the information in the biological pathways data, we represent pathways as instantiated arguments in an argumentation framework, which we present in this section.

An *abstract argument framework*, as introduced by Dung, [7] is a pair $AF = \langle \mathcal{A}, attack \rangle$, where $\mathcal{A}$ is a set of arguments and *attack* a binary relation on $\mathcal{A}$. A subset $\mathcal{B}$ of $\mathcal{A}$ is said to be *conflict-free* if no argument in $\mathcal{B}$ attacks another argument in $\mathcal{B}$. $\mathcal{B}$ is said to be *admissible* if: it is conflict-free; and it defends itself against any attack. For example, suppose arguments A_1 and A_3 are in $\mathcal{B}$, some argument A_2 is in $\mathcal{A}$ but not in $\mathcal{B}$, and A_2 attacks A_1; the set $\mathcal{B}$ is admissible when some argument in $\mathcal{B}$, such as A_3, attacks A_2. A *preferred extension* is then a maximal (with respect to set inclusion) admissible set. Several other types of extensions are defined, but they are not used in our model.

Dung's arguments are entirely abstract, with no features other than the attack relation. In order to enable some content to be given to the arguments, a refinement of Dung's abstract approach, which provides some structure for arguments, was developed in the ASPIC framework [19].

This framework assumes an unspecified logical language and knowledge base, which may include facts, strict rules, and defeasible rules; it defines arguments A_i as inference trees formed by applying inference rules (which may be either strict or defeasible) to a knowledge base: the nature of the inference rules is also unspecified, though explicitly represented in the arguments.

We represent entailment in strict rules with $\overset{s}{\rightarrow}$ and in defeasible rules with $\overset{d}{\rightarrow}$. An argument thus comprises a non-leaf node in the tree (the *conclusion* of the argument) together with the children of that node (the premises of the argument) and the rule from premises to conclusion. Leaf nodes are facts in the knowledge base. The conclusion of argument A_i can be the premise of some other argument A_j, allowing us to chain arguments together. By and large, the conclusions and premises of arguments are *literals* (atomic formulae or their negations), where we use propositional negation, e.g. if we have atomic formula p, the negation $\neg p$ is a literal. We can also have expressions of rules and their negations. It is inconsistent to have a literal or rule and its negation.

Arguments can be presented as tableau. Following [19], a strict argument, such as that labeled (A1), with premises P2 and a strict rule $[P2 \overset{s}{\rightarrow} P1]$, and conclusion P1 appears as *Strict Modus Ponens*:

$$\frac{P2, [P2 \overset{s}{\rightarrow} P1]}{P1}\ (A1)$$

While a defeasible argument (A2) with premise P4, defeasible rule $[P4 \overset{d}{\rightarrow} P5]$, and conclusion P5 appears as *Defeasible Modus Ponens*:

$$\frac{P4, [P4 \overset{d}{\rightarrow} P5]}{P5}\ (A2)$$

An argument can appear as a more extended tree, where sub-arguments of the larger argument may be strict or defeasible. A strict argument may have only sub-arguments which themselves are strict, otherwise it is a defeasible argument. For example, an argument (A3) with conclusion P11 has a defeasible intermediate argument (A4) with conclusion P9, so (A3) must then be an instance of a defeasible argumentation reasoning pattern:

$$\frac{\dfrac{[P8 \overset{d}{\rightarrow} P9], P8}{P9,}\ (A4) \qquad P10, \quad [[P9 \wedge P10] \overset{s}{\rightarrow} P11]}{P11}\ (A3)$$

As this example shows, we can have complex arguments in which strict and defeasible arguments appear as intermediate arguments.

The notion of an argument as an inference tree leads to two ways of attacking, *rebuttal* and *undermining*, an argument:[1]

[1] The literature on argumentation is more complex and diverse than presented here [19], but the simplification suits our current purposes.

- An argument A_i *rebuts* an argument A_j if the conclusion of A_i is ϕ and the conclusion of A_j is $\neg\phi$; and
- An argument A_i *undermines* an argument A_j if the conclusion of A_i is $\neg\phi$ and a premise of A_j is ϕ.

For instance, A5 rebuts A1, and A6 undermines A2. As rebuttal is a symmetric attack (unlike undercutting or undermining), we also have A1 rebuts A5:

$$\frac{P14, [P14 \overset{d}{\rightarrow} \neg P1]}{\neg P1} \; (A5)$$

$$\frac{P15, [P15 \overset{d}{\rightarrow} \neg P4]}{\neg P4} \; (A6)$$

We can abstract from the structure of the arguments and identify the AF defined by the attack relations among the arguments, for example, where AF $= \langle \{A1, A2, A3, A5, A6\}, \{att(A5, A1), att(A1, A5), att(A6, A2)\}\rangle$, then the preferred extensions are: {A3, A5, A6} and {A1, A3, A6}; nothing attacks A3 or A6, and A1 and A5 attack one another, so each extension contains one.

Concerning complexity, when an AF has no cycles, the complexity of computing the preferred extension takes time linear to the number of arguments [8]. Preliminary analysis of our working example in section 5.1 suggests that a significant portion of graphs do not produce cycles.

The strength of the approach is that we can clearly and systematically move between the instantiated and abstract arguments and their attack relations; once abstracted, the internal contents of the arguments and their specific relationships are not relevant to calculating extensions, simplifying the reasoning. Furthermore, once we have the extensions, we can then recover the content of the arguments, yielding sets of propositions which are consistent. ASPIC in [19] has a range of other components, though these are not clearly relevant for our purposes.

To this point, we have only represented propositional variables, e.g. P1,. . . , P15. However, to use the instantiated arguments, we need a knowledge base that represents the information in our domain. Typically, these appear in the form of argumentation schemes [20], which are sterotypical, defeasible reasoning patterns. There are is very large range of such patterns, including *Slippery Slope*, *Ad Homenim*, and *Practical Reasoning* [21]. We can introduce domain specific schemes such as has been proposed for legal case-based reasoning [23]. For our purpose, the knowledge base should represent schemes for *Biomolecular Pathways.* Where we represent biological information as argumentation schemes and define attacks between schemes, we can create abstract argumentation frameworks, which would allow us to reason with inconsistent data.

4 Biomolecular Action-Based Alternating Transition System

To reconstruct the pathways as argumentation graphs, we express them in a language along the lines of an Action-based Alternating Transition System (AATS) that is designed to represent *multi-agent systems* [22]. The system provides an abstract specification of sets of objects and functions. Instantiated models are defined with respect to the system, where the model satisfies axioms which are the constraints within and between components of the model as well as the incompatibilities among the objects in the components; in other words, constraints and incompatibilities provide an underlying structure to the model, which is then instantiated with particulars. In turn, argumentation schemes are instantiated with respect to the model. We present a derived structure, the *Biomolecular Action-based Alternating Transition System (BAATS).* From this, we could provide the axioms, a model, and *generate* the logical space of arguments and the attack graph over which we calculate the extensions, testing the underlying model. This is essentially the approach taken in [1].

A BAATS is a 6-tuple $S = \langle Q,\ Ac,\ \rho, \tau, \Phi, \pi \rangle$. Unlike the AATS, there are no autonomous agents, the association of actions and autonomous agents do not hold, and there are no joint actions. Propositions are associated with whether or not a biomolecule holds (in the relevant activatable context).

- Q is a finite, non-empty set of *states* $\{q_1,...,q_n\}$;
- Ac is a finite, non-empty set of interactions $\{\alpha_1,...,\alpha_n\}$;
- $\rho : Ac \rightarrow 2^Q$ is an *action pre-condition function*, which for each interaction $\alpha \in Ac$ defines the set of states $\rho(\alpha)$ from which α may be executed;
- $\tau : Q \times Ac \rightarrow Q$ is a partial *system transition function*, which defines the state $\tau(q,\ \alpha)$, that results by the performance of α from the state q, where q $\in Q$ and $\alpha \in Ac$. The function is partial as not all interactions can occur in every state;
- Φ is a finite, non-empty set of *atomic propositions*, associated with whether or not a particular biomolecule holds in the current state and location of the biological system;
- $\pi : Q \rightarrow 2^{\Phi}$ is an interpretation function, which gives the set of primitive propositions included in each state: if $p \in \pi(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q, if $\neg p$ $\in \pi(q)$, then this means that the propositional variable p is not-satisfied (equivalently, false) in state q.

Next we turn to the specification of the scheme.

5 Biomolecular Argumentation Scheme

We express a biomolecular argumentation scheme (BAS) in terms of the BAATS. In the pathways we are considering, Regulation of nuclear SMAD2/3 signaling,[2]

[2] See Pathway Interaction Database: http://pid.nci.nih.gov/

we take the positive (green arrow), neutral (black arrow), and negative (red arrow) regulators to be *premises* of a scheme connected to a *rule* (that is left implicit), which denotes the biomolecular process; the conclusion of the rule is the *result of the process given the regulators* and is most often a biomolecule. The negative (red) regulators are constraints on the application of a rule, represented in the argumentation scheme as *exceptions*, meaning that where the exception holds (i.e. true that the literal does not hold), the rule can be applied; where the exception does not hold (i.e. true that the literal does hold), the rule cannot be applied. We do not comment here on attacks on multistate biological processes. Simply put, the reactions in a pathway are recoded as propositions that have assigned *roles* (e.g. premise, exception, conclusion) in an argumentation scheme.[3]

In the following BAS, we provide an argumentation scheme in the language of the BAATS. The schemes could also be presented in tableau format, where we have defeasible rules. The premises of the scheme model the biological process and its conditions (regulators), while the conclusion represents the output of the process given that the conditions hold. The rule we leave implicit. The negative regulator is represented as an exception, which, if present, does not allow us to draw the conclusion. The main justification for the fine-grainedness of the scheme is that we want each proposition in the scheme (premises, rule, and conclusion) to represent specific 'attackable' elements. As we have more than one of each sort of regulator, we distinguish them with superscripts; the premises of each interaction are subscripted to the interaction. The premises co-occur in the same state before the application of the process.

BAS

1. α: A single step biological process.
2. ι_α: Neutral regulator with respect to α, those biomolecular elements with a neutral edge to α.
3. ϕ_α: Positive regulator with respect to α, those biomolecular elements with a positive edge to α.
4. $\neg\rho_\alpha$: Negative regulator with respect to α, those biomolecular elements with a negative edge to α.
5. $\tau(q_x,\, \alpha) = q_y$: The state transition given by α, where the presumption is that ι_α and ϕ_α regulators hold in q_x, but ρ_α does not.
6. Therefore, $\sigma \in q_y$: The result biomolecular element, where the presumption is that the element σ holds in the state q_y that results after α.

In addition to a complex scheme of premises, rule, and conclusion, we assume an *Assertion Argument*, meaning that literals can be asserted to be true without the need of premises or a rule. In Logic Programming, these are rules that have a head, but no body. In a biomedical domain, an assertion might simply be adding a drug or biomolecular to an existing process.

While we have maintained reference to a *dynamic* aspect of the processes, when it comes to evaluating the arguments in an abstract argumentation framework, we abstract over the temporal aspect to view the processes *statically*, as

[3] The image in Figure 1 is in greyscale.

co-occuring atemporally since, for the purposes of evaluation, all that matters is the *attack relation* between arguments. Yet, for the determination of *attacks* between arguments, we must consider temporality - the attacking element and the attacked element must co-occur. Formally and in an implementation, this is addressed by *unification* of variables, including a temporal variable.

5.1 An Example

For our purposes, we take the Pathway Interaction Database (PID) graph Regulation of nuclear SMAD2/3 signaling as our model which satisfies the biomolecular constraints, presuming the constraints can be defined. It is of interest since there are several negative regulators. First we provide a single instantiated argumentation scheme, then several instantiated schemes with attacks.

From the PID graph, we select the pathway in Figure 1 (modified from the PID graph for clarity). The instantiated argument is referenced as **pid_i_200106**; in the following instantiations, the prefix **pid_i** stands for *Pathway Interaction Database-Interaction-ID*, while **pid_m** stands for *Pathway Interaction Database-Molecule-ID*, where the ID is found in the graphs associated OWL file via the BioPAX link in the PID. We index premise elements with the process to distinguish between several schemes. The premise names are taken directly, without modification, from the Regulation of nuclear SMAD2/3 signaling graph.

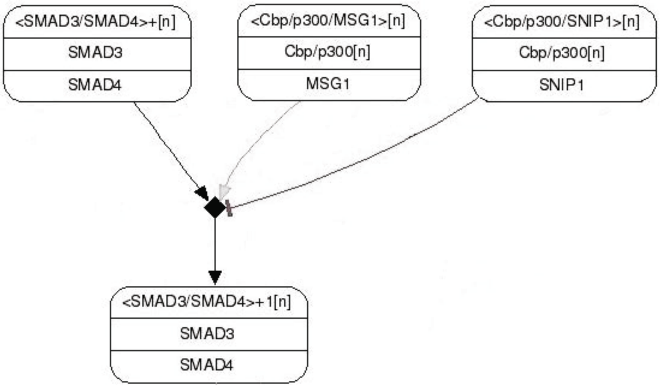


Fig. 1. Biomolecular Pathway Sample 1, where the light grey pointed arrow represents the positive regulator, the black pointed arrow represents the neutral regulator, and the flat-tipped arrow represents the negative regulator

Argument for $< SMAD3/SMAD4 > +1[n]$**; pid_i_200106**

1. α^k
2. $\iota_{\alpha^k} =< SMAD3/SMAD4 > +[n]$
3. $\phi_{\alpha^k} =< Cbp/p300/MSG1 > [n]$
4. $\neg\rho_{\alpha^k} = \neg(< Cbp/p300/SNIP1 > [n])$
5. $\tau(q_x, \alpha^k) = q_y$
6. Therefore, $< SMAD3/SMAD4 > +1[n]$ holds in q_y

To illustrate attacks between arguments, we instantiate the scheme several times (including **pid_i_200106**) using those portions of the Regulation of nuclear SMAD2/3 signaling graph that contain negative regulators.

Argument for $< SMAD3/SMAD4 > +1[n]$**; pid_i_200104**

1. α^j
2. $\iota_{\alpha^j} =< SMAD3/SMAD4 > +[n]$
3. $\phi^1_{\alpha^j} = Cbp/p300[n]$
4. $\phi^2_{\alpha^j} = PCAF[n]$
5. $\tau(q_x, \alpha^j) = q_y$
6. Therefore, $< SMAD3/SMAD4 > +1[n]$ holds in q_y

Argument for $< SMAD3/SMAD4/GR > +[n]$**; pid_i_200096**

1. α^l
2. $\iota^1_{\alpha^l} = GR$
3. $\iota^2_{\alpha^l} =< SMAD3/SMAD4 > +1[n]$
4. $\tau(q_x, \alpha^l) = q_y$
5. Therefore, $< SMAD3/SMAD4/GR > +[n]$ holds in q_y

Argument for GSC**; pid_i_200037**

1. α^m
2. $\phi_{\alpha^m} =< SMAD2/SMAD2/SMAD4/FOXH1 > +[n]$
3. $\neg\rho_{\alpha^m} = \neg(< SMAD3/SMAD4 > +1[n])$
4. $\tau(q_x, \alpha^m) = q_y$
5. Therefore, GSC holds in q_y

We can introduce assertions about molecules into the representation, e.g.:

Assertion Argument for $< Cbp/p300[n]/SNIP1 > [n]$**; pid_m_204265**

1. Therefore, $< Cbp/p300[n]/SNIP1 > [n]$ holds in q_x

5.2 An Abstract Argumentation Framework

To give some results for our system fragment, we consider the inter-relations between the instantiated schemes. Where we have undermining or rebuttal, presumptive conclusions do not follow. In particular, note that the conclusions of **pid_i_200104** and **pid_i_200106** make the *exception* of **pid_i_200037** *false*, rendering the process of **pid_i_200037** inapplicable; this means that where the processes of **pid_i_200104** and **pid_i_200106** apply, the outcome of **pid_i_200037** does not, presumably, hold as a result. It is in this way that the processes are expressed as arguments in *attack* relations. Each instantiation of a BAS can be taken as an abstract argument in an argumentation framework. And where one instantiation undermines or rebuts another instantiation, we interprete this as attack between one argument and another. For clarity, we consider three different examples of the analysis: first, we evaluate just the arguments represented in section 5.1; then, we add perturbating arguments; and finally, we make use of an *inconsistent* knowledge base and related arguments. At each point, we illustrate the impact of assumptions on the relevant extensions.

For our first example of an abstract argumentation framework, we do not consider perturbating arguments; we have the following (where we have not represented the assertion of **pid_m_204265**):

$AF_1 = \langle \mathcal{A}_1, attack_1 \rangle$, where
$\mathcal{A}_1 = \{$**pid_i_200104**, **pid_i_200106**, **pid_i_200096**, **pid_i_200037**$\}$
$attack_1 = \{<$ **pid_i_200104**, **pid_i_200037** $>$,
$<$ **pid_i_200106**, **pid_i_200037** $>\}$.

Figure 2 is a graphical representation of AF_1. Each node is an abstract representation of an instantiation of the BAS (the labels refer to the PID IDs) and the arcs between the nodes represent attacks between the arguments. For discussion below, we have introduced several arguments indicated with the variables, **x**, **y**, and **z**, that attack other arguments (as indicated). These can be read as arguments for conclusions, e.g. *an argument for x*, *an argument for y*, and *an argument for z*, and the negative conclusions.

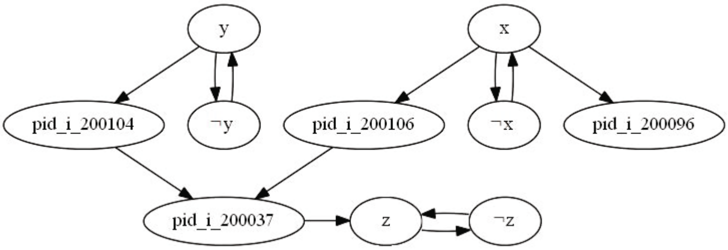


Fig. 2. Biomolecular Pathway Sample 2

Before we determine extensions, we ignore (for the moment) the variable arguments and assume that the premises of *attacking* arguments in AF_1 are all true (alternatively, have all been asserted). In this case, the preferred extension is: $PE_1 = \{$**pid_i_200104**, **pid_i_200106**, **pid_i_200096**$\}$, since **pid_i_200037** is attacked and not defended by any argument.

For our second example, we consider perturbating arguments. In our analysis, we can represent the action of particular drugs or microRNA (biological perturbating agents) on a given biological system (with instantiated arguments for **x** and **y**) as well as the effect of an outcome of a given system on other systems (with an instantiated argument for **z**). For the moment, we ignore the arguments ¬**x**, ¬**y**, and ¬**z**. We represent perturbating biological agents with instantiated BASs (and so substitute instantiated arguments for the variables **x** and **y**), where the conclusion of the argument is the particular perturbating biological agent; this conclusion is the negation of a premise (or conclusion) of some other instantiated BAS; thus, as exemplified above, one argument attacks another argument. Once instantiated, such arguments perturb the system and change its overall state, giving alternative preferred extensions, thus showing the logical effects of drugs in a biological system. In the context of our framework, the extensions capture the resulting state of the system. It is, therefore, possible to carry out *in silico* experiments which change the output of the overall process model, where different extensions are the result of a drug attack on a set of biological arguments. Alternatively, we may consider the impact of the *output* of a given biological system on *other* systems; for example, where **z** represents an argument for a desirable or at risk clinical end-point, the outcome of **pid_i_200037** may (or may not) perturb **z**, depending on the other arguments and attacks of the biological system. In this way, we can add or subtract elements, noting the overall effect on the extension and making it apparent what sorts of side-effects may arise from different combinations of elements.

To show a sample of this reasoning, we look for relevant elements from PID. Suppose **x** is the assertion associated with **pid_m_204265**, which attacks the exception premise of **pid_i_200106** and **pid_i_200096**; we presume that **pid_i_200104 is not attacked.**

$AF_2 = \langle \mathcal{A}_2, attack_2 \rangle$, where
$\mathcal{A}_2 = \{$**pid_i_200104**, **pid_i_200106**, **pid_i_200096**,
pid_i_200037, **pid_m_204265**$\}$
$attack_2 = \{< \mathbf{pid\_m\_204265}, \mathbf{pid\_i\_200106} >,$
$< \mathbf{pid\_m\_204265}, \mathbf{pid\_i\_2000096} >,$
$< \mathbf{pid\_i\_200104}, \mathbf{pid\_i\_200037} >,$
$< \mathbf{pid\_i\_200106}, \mathbf{pid\_i\_200037} >\}.$

Here, we have: $PE_2 = \{$**pid_m_204265**, **pid_i_200104**$\}$.

Alternatively, we can search for a value of **y**, which would be a biomolecule or process that attacks **pid_i_200104**, by making one of the premises false; we presume there are no attackers on **pid_i_200106** and **pid_i_200096**. For example, we can search for inhibitors of either *Cbp/p300[n]* or *PCAF[n]*, which appears to be microRNA **miR-181a/b** [24]. In such an instance, the preferred extension is $PE_3 = \{$**miR-181a/b**, **pid_i_200106**, **pid_i_200096**$\}$. When we have both **pid_m_204265** and the **miR-181a/b**, then we have $PE_4 = \{$**pid_m_204265**, **miR-181a/b**, **pid_i_200037**$\}$. By the same token, where **pid_i_200037** holds, it may serve as an attack on some process such as the variable z in Figure 2 where the conclusion of pid_i_200037 is the negation of a premise of z.

Finally, we consider all the argument nodes in Figure 2, which represents a knowledge base with inconsistent information. The arguments and attacks can be read off the graph itself. Depending on evaluations of arguments for **x**, **y**, and their negations, alternative extensions are generated as *consequences* of the attack relations. For instance, supposing ¬**y** and **x** hold, then we have $PE_5 = \{\neg\mathbf{y}, \mathbf{x}, \mathbf{pid\_i\_200104}, \mathbf{z}\}$ and $PE_6 = \{\neg\mathbf{y}, \mathbf{x}, \mathbf{pid\_i\_200104}, \neg\mathbf{z}\}$; we see that given both ¬**y** and **x**, the system is indeterminate with respect to **z**. On the other hand, where **x** and **y** holds we have a determinate result for ¬**z**: $PE_7 = \{\mathbf{x}, \mathbf{y}, \mathbf{pid\_i\_200037}, \neg\mathbf{z}\}$. Finally, where **y** and ¬**x** both hold (or where ¬**y** and ¬**x** both hold), **z** is again indeterminate.

The analysis we have presented provides some explanatory power in the sense that we can justify what appears in the extension according to what arguments are attacked or attacking, reasoning backwards through the chains of attack relations. Moreover, extracting the propositional content of the arguments, we can understand the biological terms of the explanation.

Exercising this small fragment shows how argumentation schemes and argumentation frameworks can be used to provide complex explanations for biological phenomena, to reason systematically about systems with logical inconsistencies to yield consistent sets of arguments (and the propositions they contain), which can be used to represent states of a biological system, and finally to explore the processes for their interconnections.

6 Related Work

In a series of papers, [18], [16], and [17] present an approach to and implementation of argumentation concerning biomolecular pathways. In terms of general subject area and the application of argumentation, their work and that presented here are very closely related. However, the works take different but highly complementary approaches to *what* is argued about, which is reflected in the sorts of schemes that are deployed. While we focus on an argumentation scheme most like *Practical Reasoning* in the sense that it is entirely about actions, the work of [17] focuses on *Expert Testimony*, where experts are called to present their conclusions and counter-conclusions concerning the representation of information in a database. [17] does not take statements in a database as given, but rather aim to identify contradictory statements and large consistent subsets. In our work, we assume the statements in a database as given and investigate the biological consequences of these statements. As a result, [17,18] identify "conflicting information presented by an online biological database", while we aim to identify stable modules of biological interaction networks with certain biological properties. It is a viable area of future research to investigate the relation between both levels of argumentation about biological phenomena.

Another large field of study related to our work is biological network analysis [3] and the use of biological networks in the personalized treatment of disease [4]. In each case, a crucial step is the identification of stable modules in interaction networks that are responsible for physiological processes, which correspond to pathologically abnormal functioning in the case of disease or which determine an organism's response to drugs or environmental factors. While the identification of topological and functional modules in network biology is commonly based on network clustering algorithms that break these networks down into modules of different sizes, depending on the parameters used in the clustering [3], we can identify modules based on global properties governing biological interactions. It is subject to future research to identify to which degree modules identified through our approach correspond to the modules that are traditionally identified in network biology.

7 Next Steps

The work will be developed in several directions. First, we will evaluate the accuracy of the representation and reasoning against further data; that is, are the extensions we provide consistent with experimental data and can an implementation handle data on a large scale? To assist in this evaluation over scaled up data, it will be necessary to implement a translation from databases into instantiated BASs, to determine their attack relations, and to calculate extensions. In principle, the first step is relatively straightforward, given the structure of the DB and the BAS. Furthermore, the third step already has successful implementations, where arguments and attacks are given, e.g. ASPARTIX [9]. Of more importance is the determination of *attack* relations given instantiated BASs; this can be addressed so long as strings that are used to represent biomolecules or mRNA are expressed consistently and in an appropriate literal form such that we can search for a string and its negation-prefixed form. Large scale evaluation of this work will have to wait till these issues are fully addressed. A second line of development could be to introduce some *preferential* information whereby attacks succeed or fail dependent on some additional aspect of the process. A third line of development would be to relax the assumption that attacks are, even if successful, entirely successful. In this approach, we would have attacks that introduce *degrees* of success somewhat along the lines as discussed in fuzzy logic. This might more realistically model complex biomolecular processes. Finally, it may be interesting and relevant to introduce the temporal element in reasoning in AFs.

Acknowledgements. Adam Wyner appreciates support from the FP7 IMPACT Project, grant number 247228. Luke Riley is grateful for support from the EPSRC. Robert Hoehndorf is thankful for funding from the FP7 RICORDO project, grant number 248502. The views in this paper are those of the authors.

References

1. Atkinson, K., Bench-Capon, T., Cartwright, D., Wyner, A.: Semantic models for policy deliberation. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Law, ICAIL 2011, Pittsburgh, PA, USA, pp. 81–90 (2011)
2. Atkinson, K., Bench-Capon, T.J.M.: Abstract Argumentation Scheme Frameworks. In: Dochev, D., Pistore, M., Traverso, P. (eds.) AIMSA 2008. LNCS (LNAI), vol. 5253, pp. 220–234. Springer, Heidelberg (2008)
3. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5(2), 101–113 (2004)
4. Barabási, A.L.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews. Genetics 12(1), 56–68 (2011)
5. Black, E., Atkinson, K.: Choosing persuasive arguments for action. In: Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011 (2011)
6. Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway commons, a web resource for biological pathway data. Nucleic Acids Research 39(suppl. 1), D685–D690 (2011)

7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77(2), 321–358 (1995)
8. Dunne, P., Wooldridge, M.: Complexity of abstract argumentation. In: Rahwan, I., Simari, G. (eds.) Argumentation in Artificial Intelligence, pp. 85–104. Springer (2009)
9. Egly, U., Gaggl, S.A., Woltran, S.: Answer-set programming encodings for argumentation frameworks. Argument and Computation 1(2), 147–177 (2008)
10. Hermjakob, H., Montecchi Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: Intact: an open source molecular interaction database. Nucleic Acids Research 32(suppl. 1), D452–D455 (2004)
11. Hidalgo, C.A., Blumm, N., Barabsi, A.L., Christakis, N.A.: A dynamic network approach for the study of human phenotypes. PLoS Comput. Biol. 5(4), e1000353 (2009)
12. Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature 411(6833), 41–42 (2001)
13. Kitano, H.: Computational systems biology. Nature 420(6912), 206–210 (2002)
14. Kitano, H.: Systems Biology: A Brief Overview. Science 295(5560), 1662–1664 (2002)
15. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology 6(1) (January 2010)
16. McLeod, K., Ferguson, G., Burger, A.: Using argumentation to resolve conflict in biological databases. In: Green, N., Grasso, F., Kibble, R., Reed, C. (eds.) Proceedings of Computational Models of Natural Argument (CMNA), vol. 9, pp. 15–23 (2009)
17. McLeod, K., Ferguson, G., Burger, A.: Argudas: arguing with gene expression information. In: Paschke, A., Burger, A., Splendiani, A., Marshall, M.S., Romano, P. (eds.) Proceedings of the 3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences (December 2010)
18. McLeod, K., Burger, A.: Towards the use of argumentation in bioinformatics: a gene expression case study. In: ISMB, pp. 304–312 (2008)
19. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument and Computation 1(2), 93–124 (2010)
20. Walton, D.: Argumentation Schemes for Presumptive Reasoning. Erlbaum, Mahwah (1996)
21. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press (2008)
22. Wooldridge, M., van der Hoek, W.: On obligations and normative ability: Towards a logical analysis of the social contract. Journal of Applied Logic 3(3-4), 396–420 (2005)
23. Wyner, A., Bench-Capon, T., Atkinson, K.: Formalising argumentation about legal cases. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Law, ICAIL 2011, Pittsburgh, PA, USA, pp. 1–10 (2011)
24. Zhao, J., Gong, A.Y., Zhou, R., Liu, J., Eischeid, A.N., Chen, X.M.: Downregulation of pcaf by miR-181a/b provides feedback regulation to tnv-α-induced transcription of proinflammatory genes in liver epithelial cells. The Journal of Immunology (2012)

The Use of Design Specificity in Standardized Mean Difference for Analysis of High throughput RNA Interference Screens

Karol Kozak

LMSC, ETH Zurich, Switzerland
karol.kozak@lmc.biol.ethz.ch

Abstract. RNA interference (RNAi) high-content screening (HCS) enables massive parallel gene silencing and is increasingly being used to reveal novel connections between genes and disease-relevant phenotypes. The application of genome-scale RNAi relies on the development of high quality HCS assays. Strictly standardized mean difference (SSMD), introduced by Zhang et al. [1], provides a possibility for hit selection in HCS experiments. This method has relied on normal approximation, which works in the primary screens considering positive and negative controls. This paper describes a new extension of the SSMD, which integrates bioinformatics RNAi on-target analysis results for both the *SSMD*-based testing process and the use of SSMD as a ranking metric for hit selection by using additional controls generated from RNAi libraries.

Keywords: High content screening, statistics, bioinformatics, RNAi.

1 Introduction

RNA interference (RNAi), a natural mechanism for gene silencing [2-3], has made its way as a widely used method in molecular biology in both academics and industry. Academic researchers have used RNAi to elucidate gene functions through studying a loss-of-function phenotype. Pharmaceutical and biotech companies have set up libraries for large-scale screens employing thousands of short-interfering RNA- (siRNA) or short hairpin RNA- (shRNA) encoding vectors to identify new factors involved in the molecular pathways of diseases [4-6]. RNAi has even been seen as the third class of drug targets after small molecules and proteins [7]. Based on siRNA or shRNA libraries, RNAi HCS enables massive parallel gene silencing to reveal the extent to which interference with the expression of specific genes alters the cell phenotype, and it is increasingly being used to reveal novel connections between genes and disease-relevant phenotypes [8-11].

The design of RNAi reagents is key to obtaining reliable screening results in large-scale RNAi studies. Several recent studies demonstrated that the degradation of intended transcripts by siRNA (so-called 'on-targets') and unintended effects arising from inadvertent targets (so-called 'off-targets') depend on the sequence of the RNAi reagent and have to be computational analyzed [12-19]. For knock-down/screening

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 139–150, 2012.

purposes different companies offer sets of siRNAs targeting the whole genome (or a subset of it) for various organisms. Typically, they offer at least three different siRNA, for each target gene. These siRNAs can be used either as single siRNAs or can be mixed and used as a pool of siRNAs. The main reason for offering several siRNAs per target is the varying knock-down efficiency of the individual oligonucleotide and the occurrence of off-target or non-target effects. In our study, we focus on sequence-dependent non-target siRNAs that do not bind to any mRNAs as originally designed for specific target mRNA.

Zhang et al. [20] explored statistical methods for hit selection in RNAi HCS experiments. There are two main strategies of selecting hits with large effects. One is to use certain metric(s) to rank the siRNAs by their effects and then to select the largest number of potent siRNAs that is practical for confirmation and validation assays. The other strategy is to test whether an siRNA has effects strong enough to reach a specified effect. In this strategy, we need to control the false negative and/or false-positive rates [21].

Percent viability, signal-to-noise ratio (S/N), signal-to-background ratio (S/B), SSMD and Z' factor have been used to quantify effect size of an siRNA in HCS assays. However, these metrics have issues in capturing data variability or being affected by sample size and hence cannot effectively assess the size of effect. What we are really interested in is not whether an siRNA has average activation effects being the same as the negative reference. Instead, we are interested in the false negative rate, in which the siRNAs or compounds with the large effects are not selected as hits, and the false-positive rate, in which the siRNAs with no or weak effects are selected as hits, in the process of hit selection. To address this question, Zhang [20] proposed an SSMD-based process with a flexible and balanced control of false positives and false negatives in hit selection.

Here, we illustrate the shortcomings of a statistical parameter that measures the magnitude of both paired and unpaired differences and thus can be used to measure the magnitude of impact of siRNAs in both primary and confirmatory screens. For the hit selection analysis, we propose a SSMD considering non-target siRNAs to the measured intensity of each siRNA individual. Currently presented SSMD is based on two types of controls, which may not be optimal for currently available RNAi libraries. This paper proposes an algorithm which extends existing SSMD parameter to deal with problem of siRNA design quality by using on-target analysis.

2 Results

2.1 SSMD

A recently proposed parameter SSMD [22], measures the magnitude of impact more effectively than any other currently used metrics. SSMD has been applied for quality control in genome scale RNAi research [22-24]. Utilizing the fact that SSMD effectively measures the size of effect, Zhang proposes an SSMD-based hit selection method to maintain a balanced control [25]. This method has also been applied to select hits in RNAi HCS primary experiments [26].

Consider two independent populations (eg. positive or negative controls), A_1 and A_2, and let D be the difference between the two populations. SSMD (β) is defined as the ratio of the mean to the standard deviation (sd) of D—namely,

$$\beta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \qquad (1)$$

μ_1 = Mean of A_1

μ_2 = Mean of A_2

σ_1 = Standard deviation of A_1

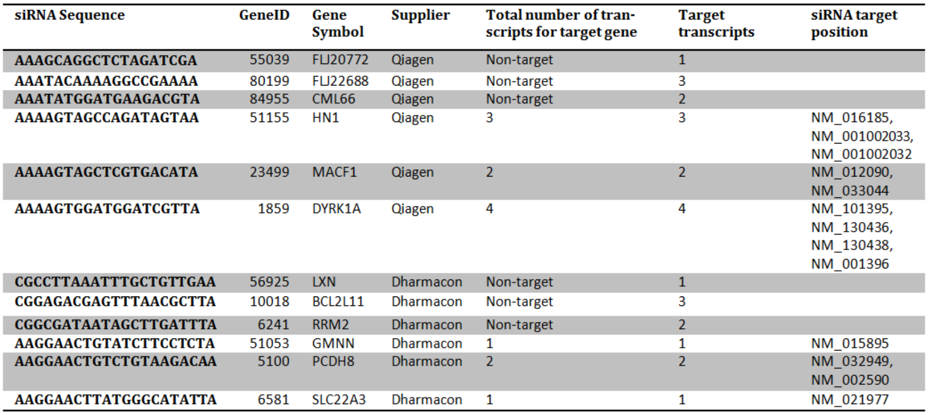
σ_2 = Standard deviation of A_2

SSMD measures the strength of the difference between two compared groups and is a good candidate for assessing differences between an siRNA and negative controls [22, 25]. For instance, when a large proportion of the sample wells are expected to be negative, we might exclude the largest 1% values and the smallest 1% values and use the remaining 98% values in the sample wells as a negative reference group in each plate to calculate percentage inhibition, estimated SSMD value, and z-score in primary screens [25]. If there are difficulties with the negative controls (e.g., they are small in number), then we recommend to use some of non-targets siRNA from sample wells to estimate a negative reference distribution [27].

2.2 On-target Analysis and Non-target siRNA

The specificity of the siRNA sequence is a crucial factor in an RNAi experiment [28]. Gene expression silencing through the RNAi machinery works perfectly if the siRNA is totally complementarity to its target mRNA. Single nucleotide mismatches between the siRNA and the target mRNA decrease the rate of mRNA degradation [29, 30]. The algorithms of the different companies for generating the best siRNA sequence typically take this into account and check and exclude siRNA sequences which have total complementarity to other than the target mRNA. Nevertheless, with evolution of RefSeq database [31], already designed siRNAs, today may not match to any mRNA sequence. In our study, we focus on sequence-dependent analysis which identify non-target constructs. On-target analysis provides the latest reagent annotations and internally calculates quality information, such as predicted specificity and efficiency. SiRNA design with old version RefSeq may have no homology to any known mammalian gene in latest RefSeq. Such siRNA can be treated as well as negative control. Sequence dependent analysis of 2 genome wide libraries was evaluated for non-target specificity by performing a search against the latest RefSeq (09.2011) database (see methods section). Analysis of different sequence dependent parameters is summarized in discussion section. Minimal nonspecific effects ensure that comparison of the gene-specific siRNA to the non-target siRNA or negative control gives a true picture of the effects of target-gene knockdown on gene expression and phenotype. If the non-targets or negative control causes nonspecific effects then results from RNAi experiments can be misleading and difficult to interpret. There are advantages of using non-targets as additional negative controls to validate a quality of screening assay: I. Results from non-target siRNA can be compared to results from untransfected cells to determine whether the experimental setup causes nonspecific effects. II. Results from non-target siRNA can be compared to results from gene-specific siRNA to pinpoint the effects of target gene knockdown. III. A larger number of negative controls will provide adequate estimation of their mean. If altered expression or phenotype are observed in cells transfected with non-target siRNA, these changes are nonspecific, i.e., due to transfection procedures or siRNA toxicity and not sequence complementarity. Nonspecific effects should be minimal to ensure reliable RNAi results. Results from the non-targets or negative controls can also be compared to results from the gene-specific siRNA under study.

Table 1. Selected on-target and non-target siRNAs from Qiagen and Dharmacon library analyzed with RefSeq 2011



siRNA Sequence	GeneID	Gene Symbol	Supplier	Total number of transcripts for target gene	Target transcripts	siRNA target position
AAAGCAGGCTCTAGATCGA	55039	FLJ20772	Qiagen	Non-target	1	
AAATACAAAAGGCCGAAAA	80199	FLJ22688	Qiagen	Non-target	3	
AAATATGGATGAAGACGTA	84955	CML66	Qiagen	Non-target	2	
AAAAGTAGCCAGATAGTAA	51155	HN1	Qiagen	3	3	NM_016185, NM_001002033, NM_001002032
AAAAGTAGCTCGTGACATA	23499	MACF1	Qiagen	2	2	NM_012090, NM_033044
AAAAGTGGATGGATCGTTA	1859	DYRK1A	Qiagen	4	4	NM_101395, NM_130436, NM_130438, NM_001396
CGCCTTAAATTTGCTGTTGAA	56925	LXN	Dharmacon	Non-target	1	
CGGAGACGAGTTTAACGCTTA	10018	BCL2L11	Dharmacon	Non-target	3	
CGGCGATAATAGCTTGATTTA	6241	RRM2	Dharmacon	Non-target	2	
AAGGAACTGTATCTTCCTCTA	51053	GMNN	Dharmacon	1	1	NM_015895
AAGGAACTGTCTGTAAGACAA	5100	PCDH8	Dharmacon	2	2	NM_032949, NM_002590
AAGGAACTTATGGGCATATTA	6581	SLC22A3	Dharmacon	1	1	NM_021977

To detect non-targets using the sequence information, all RNAi constructs were computationally mapped onto the latest genomic sequence RefSeq using homology search algorithm. Annotations for targeted genes and transcripts were derived through the mapping on genome and transcriptome databases. siRNAs which had minimum alignment score 95 and length of full match 19 were retained in the database as on-target siRNA. This pre-computed database provides a list of siRNAs with highest possible score (greatest knockdown). This enabled the selection of 71593 siRNAs in case of Qiagen library, 71764 siRNAs in case of Dharmacon library. All remaining siRNA were retained in the database and assign as non-target siRNA (Fig. 1). This allowed us in case of Qiagen library to select 1178 siRNAs, in case of Dharmacon library to select 794 siRNAs. On an average approximately 2% siRNAs per library are labeled as non-target reagents. The Table 1 is demonstrating a list of selected siRNA (on-target and non-target), their sequences and on-target analysis results.

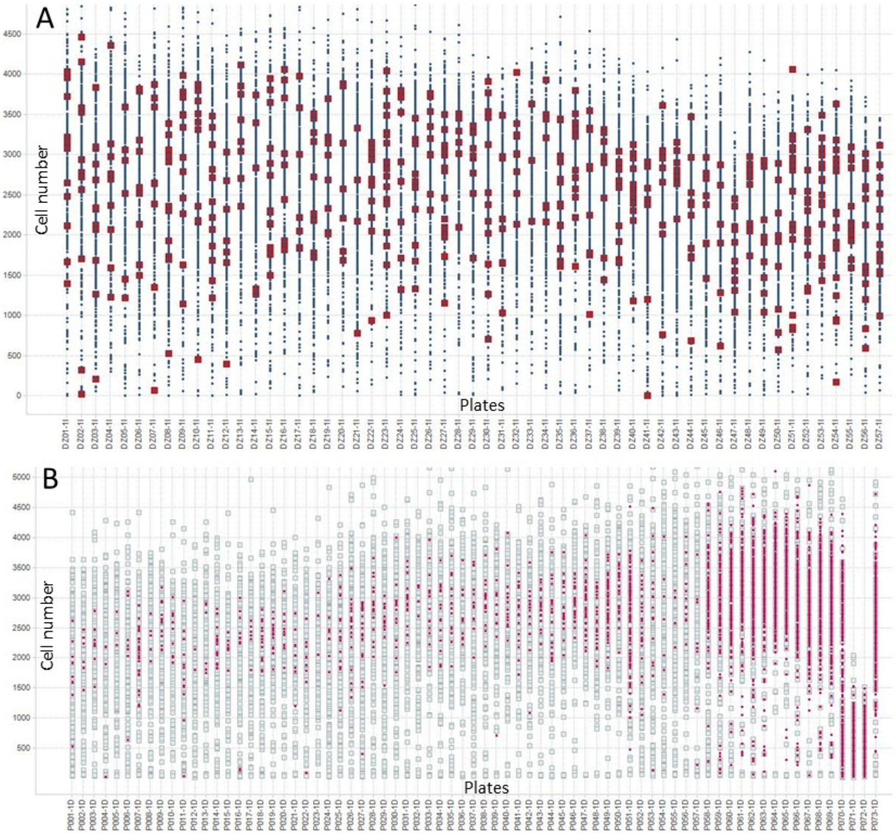


Fig. 1. Distribution of non-target siRNA on A) 57 of 384 well plates for genome wide Dharmacon library including 4 pooled oligonucleotides B) 73 of 384 well plates for genome wide Qiagen library including 4 pooled oligonucleotides. Mapping analysis has been made based on transcriptomes from RefSeq 2011 [31].

2.3 Non-target SSMD

SSMD accurately captured the clear difference between the high and low populations. For example, cost and convenience aside, would better results be obtained from using 1 low and high control per experimental unit or 10 per unit? Quite clearly, the greater the number of controls, the better the precision of the results is concerned. Should 1 control or 6 or 26 be used? In practice, these questions are usually answered on the basis of cost, throughput, assay format, logistics, automation capabilities, and so forth. Extending amount of controls with additional wells including non-target reagents can improve precision of hit selection results without changing biological assay. As a summary of the preceding discussion, the non-target SSMD is given below:

$$\beta = \frac{\mu_1 - (\mu_2 + \mu_n)}{\sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_n^2}} \tag{2}$$

μ_1 = Mean of A_1
μ_2 = Mean of A_2
μ_n = Mean of non-target siRNAs
σ_1 = Standard deviation of A_1
σ_2 = Standard deviation of A_2
σ_n = Standard deviation of non-target siRNAs

To validate non-target SSMD in this paper we concentrate on multiwell plates extracted from RNAi HCS experiments, which may have different data ranges, different numbers of positive and negative control wells and different distribution of non-target siRNAs (Fig. 2), so that we can see the impact of sample size, siRNA design and data range on the results metrics. The biogenesis project (4 siRNA/target gene individually distributed - siRNA extension of kinome screen [32]) is dealing with ribosomes, which are macromolecular complexes used to synthesize proteins.

For calculation purposes, 4 siRNA per target gene were computationally pooled in 75x384 well plates.

Average z' score and non-target SSMD values for this screen are represented on Fig. 2. Plots on Fig. 2 shows the raw intensity (SSMD, Z' score), where we can see that the positive control wells have apparently higher intensities in plates 62 to 75 than in the remaining plates. On the other hand, both negative control and non-targets in plates 8 to 16 have intensities apparently higher than those in the remaining plates.

Non-target SSMD in the sample wells are fairly stable, more stable than Z' score across plates (Fig. 2, A, B), and they are not affected by the errors in the positive control wells. It is notable that the shape of non-target SSMD values is the same as that of z-score values in Fig. 2 A, B. Actually, for the siRNAs without any replicate, the SSMD value has a linear relationship to the corresponding z' score value—namely if both are calculated using the same negative reference. Therefore, in terms of ranking siRNA effects, the SSMD method is equivalent to the z-score method, although the z' score value is times the estimated non-target SSMD value.

If, as an example, one would decide to select 500 siRNAs from an HCS experiment for further confirmatory experiment, the specific siRNAs selected will be highly dependent on the selection methods employed. Fig. 3 shows a comparison between hit ranking based on SSMD relative to the negative reference and hit selection based on non-target SSMD (including non-targets for negative reference). Selecting the 500 siRNAs with the largest non-target SSMD and graphing the number of potential hits on a plate-by-plate basis, it is apparent that there is a hill in plates 30, 46, 54, 67 and a clear valley in plates 2, 18. Among these 500 hits, there were 6.6 potential hits per plate in the 75 normal plates on average. The number of potential hits per plate in plates 47, 56 was about 2 times that in the normal plates.

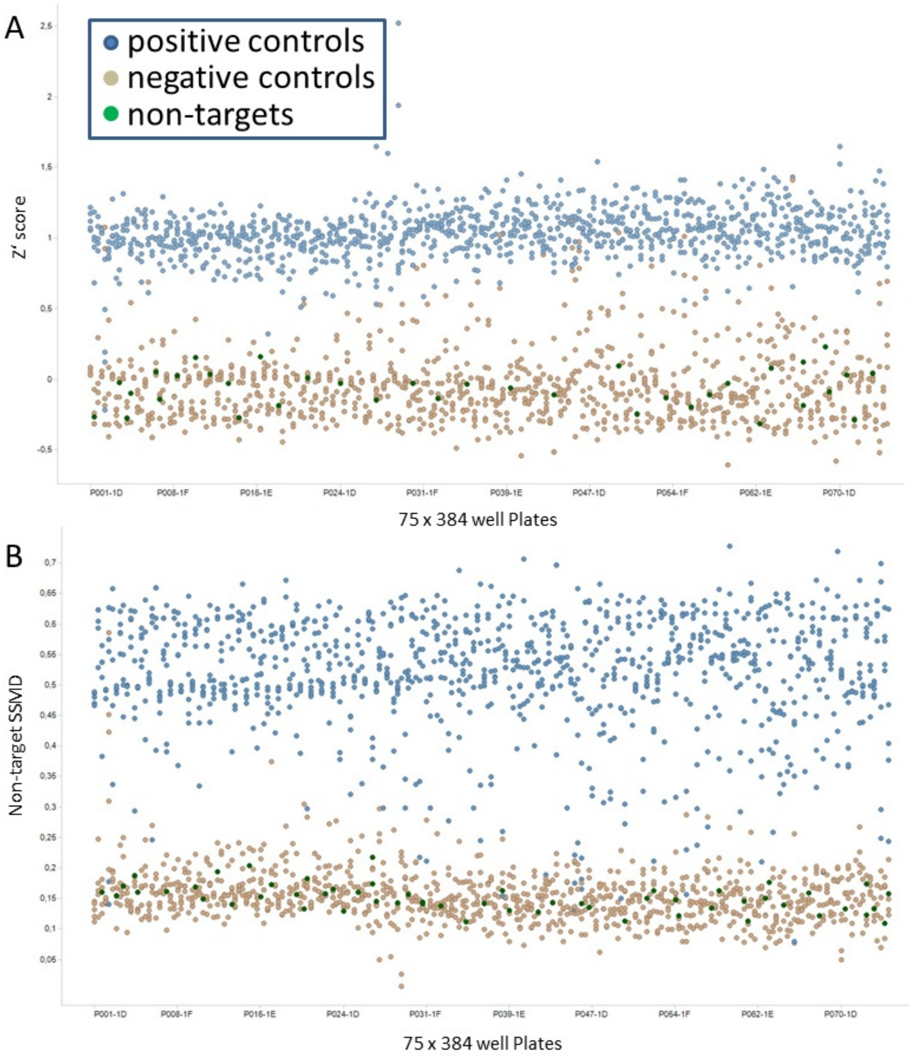


Fig. 2. Plots of 75 Plates from Genome wide experiment to display *z'* score and non-target *SSMD* for ranking siRNA hits in biogenesis primary data. Panels **A**, **B** display raw intensity in –log2 scale, estimated value of z' score and non-target *SSMD* value. In each panel, a point denotes a control and non-target wells, and the plate numbers are labeled in the *x*-axis; blue, brawn, green represent the positive controls, negative control, and non-target control wells, respectively.

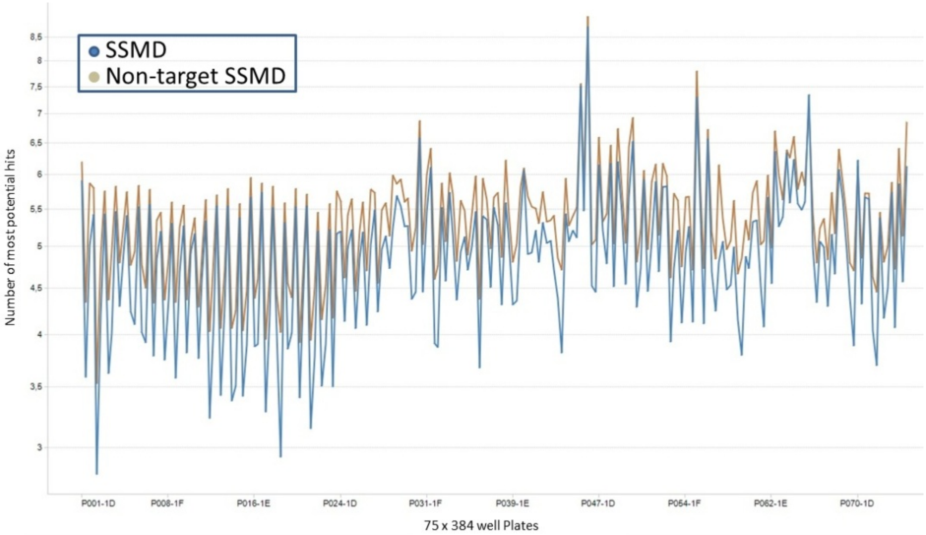


Fig. 3. The number of potential hits per plate using ranking strategy: green line- the estimated strictly SSMD values of positive and negative controls as a group, brown line - the estimated non-target *SSMD* values of positive, negative controls and non-targets as a group.

3 Methods

In this project the biogenesis of the small ribosomal subunit (40S subunit) has been used for comparison of new designed SSMD method. Image analysis in biogenesis screen was performed in an automated fashion, using image segmentation based on the identification of cell nuclei by Hoechst fluorescence as a first step. For example, nucleolar and nucleoplasmic accumulation of Rps2-YFP were distinguished by textural features. By performing a genome-wide siRNA screen following parameters have been used: 2 classes positive/negative, 27 image features, 500 cells per well, 3 channels, 4 oligo, total number of observations (records,-rows) = 108324. Image segmentation parameters: 1: green mean intensity nuclei, 2: green std intensity nuclei, 3: green mean intensity cytoplasm, 4: green std intensity cytoplasm, 5: green mean intensity cells, 6: green std intensity cells, 7: blue mean intensity nuclei, 8: blue std intensity nuclei, 9: blue mean intensity cytoplasm, 10: blue std intensity cytoplasm, 11: blue mean intensity cells, 12: blue std intensity cells, 13-27: nuclei texture green. The use of a specific assay enables the visual detection of nuclear 40S maturation defects upon depletion of a protein by RNAi. In total: for Qiagen library 17 632 genes are targeted by four different oligo (total number of oligo 72771). All calculations for on-target analysis were performed using a new design/evaluation pipeline, HCDC-KNIME [33]. In the first step, the data needed for the non-target analysis is supplied. Since the analysis is based on sequence-dependent interactions between siRNA and mRNA, the sequence information for every siRNA has to be given, which usually is provided by the companies, which designed the siRNAs. The resulting list of a complementarity search can be too long to find the important results just by visual inspection. Therefore the next step is to filter this list to reduce its size to meaningful results. Complementarity search: In this analysis step, it can be determined if there exists a complementary region between the selected siRNA sequences and the mRNAs. Many different sequence alignment algorithms to perform such a complementarity search are available, but they are not optimal for the purpose of this process step. Therefore, three different strategies for the use of these algorithms have been developed to find nearly exact complementary regions as well as small local complementarities. BLAST search: The Basic Local Alignment Search Tool (BLAST) [34] is one of the most popular algorithms for complementarity search and can be applied to find nearly identical gene regions for a specific siRNA sequence. BLAST is an effective tool to find out immediately if obvious on-target siRNA exist with a full identical nucleotide sequence to the mRNA. Therefore the BLAST search against the mRNA database from the RefSeq project is the first strategy for a complementarity search in this concept. In contrast to BLAST, other local alignment algorithms can find small partial complementarities between siRNA and mRNA sequences. Therefore, the developed concept offers, besides the BLAST search, two different alternatives of building a local alignment without getting into the mentioned runtime problem. Smith-Waterman algorithm: The Smith-Waterman algorithm is an accurate algorithm used to build local alignments between two sequences [35]. Since its use with all mRNAs from the database is not practicable, a feasible alternative is to limit the number of mRNAs to approximately 200 of mRNAs from result list. Seed-Motif-Search combined with the Smith-Waterman algorithm: Because of the mentioned runtime problem when performing a local alignment with the Smith-Waterman algorithm, a third variant to search for complementarity is introduced here. In this variant, an initial step reduces the length of the mRNA sequences to enable the use of a local alignment algorithm. This reduction is made because the seed region of the siRNA seems to play a significant role in causing non-target effects. At the beginning, all occurrences of the seed motif of every siRNA are localized in the genes (see Fig. 4). After detecting this small region, a sequence of ~50 nt around this seed motif is cut out in the mRNA. Thus, as a result of this first step, a huge number of sequences of ~50 nt in length are obtained containing the seed region of each siRNA. Due to the small length of the sequences it is now possible to perform a local alignment with the Smith-Waterman algorithm. The advantage of the Seed-Motif-Search for non-target analysis is that it limits the results to those genes which perfectly match with the seed region of the siRNA.

4 Discussion

The statistical method outlined here, while certainly not comprehensive, offer effective, practical tools for RNAi screening, using siRNA libraries updated with latest genome annotation. We have applied the on-target algorithms to re-annotate libraries and produce list of non-target reagents used later in SSMD as additional controls. Our analysis of two genome-wide RNAi libraries for Human revealed differences in genome coverage and predicted quality (for example, off-targets, on-targets). The differences most likely depending on two factors: the quality of the underlying genome release and the factors known to influence reagent quality at the time of the library design. One is to use SSMD method similar to signal-to-noise ratio, percentage inhibition, or p-value from either the z-score method or t-test of testing mean difference, to rank the siRNAs by potency and then to select a specified number of the strongest siRNAs for further investigation. In addition, several recent studies demonstrated that the degradation of intended transcripts by siRNA (so-called 'on-targets') and lack of degradation (described 'non-targets') depend on the sequence of the RNAi reagent and have to be computational analyzed. Beside non-targets there are other not effective siRNA candidates in libraries which can be treated as negative controls. For example, A 4 identical bases in siRNA sequence [GGGG, CCCC, TTTT, AAAA] can cause potential nonspecific effects through its interaction with heparin-binding proteins. In addition siRNA sequences shouldn't contain specific motifs, 5'-TGTGT-3' or 5'-GTCCTTCAA-3' in the guide strand of siRNA duplexes. We have exploited in existing libraries about 2% of siRNAs from Qiagen, 1.8% of siRNAs from Dharmacon having in sequence 5'-TGTGT-3' motifs. This paper presented a new statistical scoring method non-target SSMD for HCS considering sequence dependent effects in reagents that were designed to deal with low number of negative controls on screening plates. However, if there are no non-targets on plate, then traditional SSMD has to be applied for quantifying siRNA effects.

References

1. Zhang, X.H.D.: Genome-wide screens for effective siRNAs through assessing the size of siRNA effects. BMC Research Notes 1186(1), 33 (2008)
2. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C.: Potent and spe-cific genetic interference by double-stranded RNA in Caenorhabditiselegans. Nature 391, 806–811 (1998)
3. Hannon, G.J.: RNA: A Guide to Gene Silencing. Cold Spring Harbor Laboratory Press, New York (2003)
4. Kurreck, J.: RNA interference: perspectives and caveats. RNA Interference Gene Silencing 1, 50–51 (2005)
5. Mahanthappa, N.: Translating RNA interference into therapies for human diseases. Pharmacogenomics 6, 879–883 (2005)
6. Whelan, J.: First clinical data on RNAi. Drug Discovery Today 10, 1014–1015 (2005)
7. Nature News, Silent running: the race to the clinic. Nature, 442, 614–615 (2006)
8. Zuck, P., Murray, E.M., Stec, E., Grobler, J.A., Simon, A.J., Strulovici, B., et al.: A cell-based β-lactamase reporter gene assay for the identification of inhibitors of hepatitis C virus replication. Anal. Biochem. 334, 344–355 (2004)
9. MacKeigan, J.P., Murphy, L.O., Blenis, J.: Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. Nat. Cell. Biol. 7, 591–600 (2005)
10. Nybakken, K., Vokes, S., Lin, T.Y., McMahon, A.P., Perrimon, N.A.: Genome-wide RNA in-terference screen in Drosophila melanogaster cells for new components of the HH signal-ling pathway. Nat. Genet. 37, 1323–1332 (2005)
11. Pelkmans, L., Fava, E., Grabner, H., Hannus, M., Habermann, B., Krausz, E., et al.: Genome-wide analysis of human kinases in clathrinandcaveolae/raft-mediated endocytosis. Nature 436, 78–86 (2005)

12. Amarzguioui, M., Prydz, H.: An algorithm for selection of functional siRNA sequences. Biochem. Biophys. Res. Commun. 316, 1050–1058 (2004)
13. Chiu, Y.L., Rana, T.M.: RNAi in human cells: basic structural and functional features of small interfering RNA. Mol. Cell. 10, 549–561 (2002)
14. Khvorova, A., Reynolds, A., Jayasena, S.D.: Functional siRNAs and miRNAs exhibit strand bias. Cell. 115, 209–216 (2003)
15. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., Khvorova, A.: Rational siRNA design for RNA interference. Nat. Biotechnol. 22, 326–330 (2004)
16. Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., Zamore, P.D.: Asymmetry in the as-sembly of the RNAi enzyme complex. Cell. 115, 199–208 (2003)
17. Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., Saigo, K.: Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Res. 32, 936–948 (2004)
18. Shah, J.K., Garner, H.R., White, M.A., Shames, D.S., Minna, J.D.: sIR: siRNA Information Resource, a web-based tool for siRNA sequence design and analysis and an open access siRNA database. BMC Bioinformatics 8, 178 (2007)
19. Wang, X., Varma, R.K., Beauchamp, L., Magdaleno, S., Sendera, T.J.: Selection of hyperfunctionalsiRNAs with improved potency and specificity. Nucleic Acids Res. 37, e152 (2009)
20. Zhang, J.H., Chung, T.D.Y., Oldenburg, K.: A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J. Biomol. Screen. 4, 67–73 (1999)
21. Zhang, X.H.D., Lacson, R., Yang, R., Marine, S.D., McCampbell, T.D.M., Hare, T.R., Kajdas, J., Berger, J.P., Holder, D.J., Heyse, J.F., Ferrer, M.: The use of SSMD-based false discovery and false non-discovery rates in genome-scale RNAi screens. Journal of Biomolecular Screening 15, 1123–1131 (2010)
22. Zhang, X.D.: A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. Genomics 89, 552–561 (2007)
23. Zhang, X.D.: Novel analytic criteria and effective plate designs for quality control in genome-wide RNAi screens. Journal of Biomolecular Screening 13, 363–377 (2008)
24. Zhang, X.D., Espeseth, A.S., Johnson, E.N., Chin, J., Gates, A., Mitnaul, L.J., Marine, S.D., Tian, J., Stec, E.M., Kunapuli, P., Holder, D.J., Heyse, J.F., Strulovici, B., Ferrer, M.: Integrating experimental and analytic approaches to improve data quality in genome-wide screens. Journal of Biomolecular Screening 13, 378–389 (2008)
25. Zhang, X.D.: A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. Journal of Biomolecular Screening 12, 645–655 (2007)
26. Zhang, X.D., Ferrer, M., Espeseth, A.S., Marine, S.D., Stec, E.M., Crackower, M.A., Holder, D.J., Heyse, J.F., Strulovici, B.: The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments. Journal of Biomolecular Screening 12, 497–509 (2007)
27. Mazur, S., Csucs, G., Kozak, K.: Z' Factor including siRNA design quality parameter in RNAi screening experiments. RNA Biology (in press, 2012)
28. Semizarov, D., Frost, L., Sarthy, A., Kroeger, P.A., Halbert, D.N., Fesik, S.W.: Specificity of short interfering rna determined through gene expression signatures. Proc. Natl. Acad. Sci. USA 100(11), 6347–6352 (2003)
29. Haley, B., Zamore, P.D.: Kinetic analysis of the rnai enzyme complex. Nat. Struct. Mol. Biol. 11, 599–606 (2004)

30. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., Tuschl, T.: Functional anat-omy of sirnas for mediating efficient rnai in drosophila melanogaster embryo lysate. EMBO Journal 20, 6877–6888 (2001)
31. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65 (2007)
32. Wild, T., Horvath, P., Wyler, E., Widmann, B., Badertscher, L., Zemp, I., Kozak, K., Csusc, G., Lund, E., Kutay, U.: A protein inventory of human ribosome biogenesis reveals an essential function of Exportin 5 in 60S subunit export. PLoS. Biol. 8, e1000522 (2010)
33. HCDC web page, bioinformatics module, `http://hcdc.ethz.ch/index.php?view=article&id=25`
34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)

Toward a Translational Medicine Approach for Hypertrophic Cardiomyopathy

Catia M. Machado[1], Francisco M. Couto[1], Alexandra R. Fernandes[2,3,4], Susana Santos[2,3], and Ana T. Freitas[5]

[1] LaSIGE, Departamento de Informática, Universidade de Lisboa, Lisboa, Portugal
cmachado@xldb.di.fc.ul.pt
[2] Universidade Lusófona de Humanidades e Tecnologias, Lisboa, Portugal
[3] Centro de Química Estrutural, Instituto Superior Técnico, Lisboa, Portugal
[4] Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
[5] Instituto de Engenharia de Sistemas e Computadores, Instituto Superior Técnico, Lisboa, Portugal

Abstract. Hypertrophic cardiomyopathy (HCM) is a complex genetic disease characterized by a variable clinical presentation and onset, as well as a high number of associated mutations.

Therefore, this disease is a good candidate for a translational medicine approach to assist in its prognosis. For this purpose, we propose a framework containing two components: one for data integration, and another for data analysis based on clinical-genetic associations obtained with data mining techniques.

In this article we present the implementation of the first component. At its basis is a semantic data model developed in OWL representing the clinical and genetic data necessary for the characterization of HCM patients. This model follows a modular approach and includes mappings to controlled vocabularies such as the NCI Thesaurus and SNOMED-Clinical Terms.

The development of the model has been done in collaboration with biomedical experts, who are also the providers of the data to populate it.

Keywords: translational medicine, data integration, data mining, hypertrophic cardiomyopathy, clinical decision support systems.

1 Introduction

Hypertrophic cardiomyopathy (HCM) is a genetic disease that may afflict as many as 1 in 500 individuals, and is the most frequent cause of sudden cardiac death among apparently healthy young people and athletes [1,2]. It is characterized by a variable clinical presentation and onset, which results in a difficult clinical diagnosis prior to the development of severe or even fatal symptoms [1,2]. Moreover, its genetic diagnosis is complex, since there are approximately

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 151–165, 2012.

900 mutations in more than 30 genes currently known to be associated with the disease [3].

In terms of prognosis, the task is by no means trivial since the severity of HCM varies even between direct relatives. It has been observed that the presence of a given mutation can correspond to a benign manifestation in one individual and result in sudden cardiac death in another [1,2].

As a consequence of all these factors, HCM is an example of a disease that can benefit from a translational medicine approach to aid in its prognostic. Given the clinical manifestations of the disease and the mutations associated with it, it might be possible to identify a set of factors that will aid cardiologists in the task of risk assessment and management. This task is of paramount importance as it could enable the timely identification of patients prone to sudden cardiac death, and the regulation of their physical activities in order to minimize such risk.

A pivotal step toward the concretization of a translational medicine approach consists in the integration of data originating from different domains of knowledge. Ontologies, and controlled vocabularies in general, are important tools for data integration since they provide a standard way of representing knowledge. Ideally, these vocabularies are references accepted by the community, such as the Gene Ontology [4] and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [5].

Due to their importance in data integration, ontologies are a central piece in the implementation of the Semantic Web vision proposed by Tim Berners-Lee [6]. This vision is that of a Web of data, rather than the Web of documents that is the current standard. The general idea is that instead of links connecting Internet pages that are mostly designed to be interpreted by humans, we can have links connecting the data elements themselves. The Semantic Web can be seen as a framework for data integration at a Web-wide scale that is independent of the domain of knowledge, and focused on the meaning and on the context of the data.

In order to implement this vision of a new Web, a set of tools and technologies have been proposed as standards by the World Wide Web Consortium (W3C) [7], namely: RDF, a language for data representation and interchange [8]; and OWL, a language to formally define meaning in Web resources that supports reasoning [9].

Several translational medicine examples exist using Semantic Web standard technologies, namely the work developed by Gudivada and colleagues [10] in a task of gene prioritization, ASSIST (Association Studies aSsisted by Inference and Semantic Technologies) [11], and the Neuroweb system [12]. In the first example, the integrated resources are locally maintained in relational format and instantly converted to RDF upon need, whereas in the other two they are maintained in their original format and location. In all three examples, data integration is mediated by an ontology developed in OWL. None of the systems reuse existing ontologies, although the developers of Neuroweb considered the use of resources such as SNOMED-CT and the Disease Ontology. However, the authors verified that SNOMED-CT did not provide a suitable formulation of concepts for their purpose, and that the taxonomy adopted by the Disease Ontology was different from the one used by the clinicians participating in the Neuroweb network.

In our translational medicine approach for HCM, we are interested in the prognosis of the disease. Our goal is the identification of associations between clinical and genetic factors that can be used to aid medical doctors in the prediction of the outcome of the disease, for every individual patient, particularly in respect to the occurrence of sudden cardiac death. For this purpose, we propose a framework that integrates clinical and genetic data mediated by a semantic data model representing the disease, and that explores data mining models depicting the clinical-genetic associations (Figure 1).

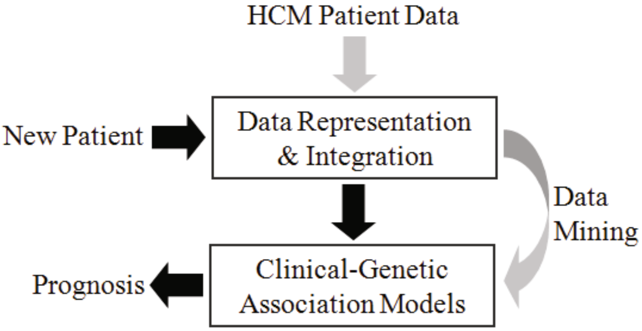

HCM Patient Data
New Patient
Data Representation & Integration
Data Mining
Clinical-Genetic Association Models
Prognosis

Fig. 1. High-level schematic representation of the translational medicine framework we are developing for the disease HCM. Data of patients with known prognosis (*grey arrows*) will be represented and integrated according to a semantic data model developed for the disease, and will be explored with data mining techniques to obtain clinical-genetic association models. Data of new patients, with unknown prognosis (*black arrows*), will be represented according to the semantic model, and will be evaluated based on the clinical-genetic association models to obtain a prediction of the prognosis.

In this article we present the implementation of the data representation and integration component. The semantic data model at the core of this component provides a useful framework for the integration of data from two different domains of knowledge, clinical and genetic, and from different institutions. The concepts modeled were identified and defined with the help of medical doctors, geneticists and molecular biologists based on the data elements collected during their activities. The model is currently being populated with data from four medical institutions and two research centers.

The rest of this document is organized as follows: Section 2 describes the development of the semantic model; Section 3 presents the semantic model; Sections 4 and 5 contain a Discussion and the Conclusions, respectively.

2 Semantic Model Development

The development of the HCM semantic model followed the guidelines presented by Noy and McGuinness [13] for ontology development. The first step was the definition of the domain (i.e. the disease, HCM) and the scope (i.e. the representation of the data necessary for the diagnosis and the prognosis of HCM), followed by the enumeration of relevant concepts and the reuse of existing controlled vocabularies.

The following steps describe our approach:

1. An initial set of concepts was identified in collaboration with biomedical experts.
2. The concepts were represented in OWL Lite, including hierarchical and non-hierarchical relations.
3. Existing controlled vocabularies of interest were searched.
4. New concepts to consider were identified in these controlled vocabularies.
5. The concepts and relations represented were continuously validated by the biomedical experts.
6. The consistency of the model (i.e. the absence of syntactic or semantic errors) was evaluated periodically.

The model was developed in the Protégé-OWL editor (version 3.4.2) [14] following a modular approach. The consistency evaluations were performed by running the reasoner HermiT [15] available in Protégé.

OWL was the language of choice to comply with the Semantic Web standards and to take advantage of external resources published in the Semantic Web.

The identification of controlled vocabularies of interested was performed using BioPortal, from the National Center for Biomedical Ontology [16]. The concepts initially identified in collaboration with the biomedical experts were used as search terms, namely *clinical history*, *angina*, *hypertrophic cardiomyopathy*, *resuscitated sudden death*, and *electrocardiography*. We searched for vocabularies referring to the medical and molecular biology domains that contained the concepts of interest, and that represented these concepts in a hierarchical organization in accordance with the vision of the HCM domain conveyed by the experts. The adequacy of the vocabularies was evaluated based on their scope. The list initially compiled was narrowed down based on the number of concepts of interest the vocabulary contained.

As previously published [17], three vocabularies were initially identified and considered for the HCM model: SNOMED CT (version 2010_01_31), the National Cancer Institute Thesaurus (NCIt) (version 10.03)[18], and the Ontology of Clinical Research (OCRe) (version 0.95) [19].

We opted to use more than one vocabulary for each module for two reasons:

(*i*) none of the vocabularies contained a complete list of the concepts of interest; (*ii*) the provided representation of the concepts was not always the most suitable for our purposes.

We did not reuse entire modules of any of the vocabularies since our goal was not to convey the most complete representation of the disease. We rather wanted to represent the concepts necessary for its diagnosis and prognosis, as well as include a minimum set of concepts that would facilitate the mapping between the HCM model and the vocabularies. In addition, one of the concerns during the development of the model was to maintain it as simple as possible, in order to avoid overwhelming the biomedical end-users with superfluous information.

Our approach to the use of these vocabularies is summarized in the following steps:

1. The regions of interest in each vocabulary were identified.
2. The hierarchical structure of the HCM model was refined in accordance with the vocabulary considered.
3. The concepts in the model were renamed in accordance to the vocabulary.
4. The concepts in the model were manually mapped to the equivalent concept in the controlled vocabulary, through a *hasDbXRef* property [1].
5. When the vocabulary provided a definition for the mapped concept, it was added to the model.

Considering that the controlled vocabularies were also exploited to identify new concepts to include in the model, they served the dual purpose of aiding in the development of the model and providing mappings.

Since its preliminary version [17], the HCM model has been extended in number of concepts and mappings. One of the previously considered vocabularies, OCRe, was eliminated due to the deprecation of the concepts we had reused (e.g. *Health Care Site*). The model is currently mapped to four controlled vocabularies: SNOMED CT and the NCIt as before, and also to the Gene Regulation Ontology (version 0.5, released on 04_20_2010) [20] and to the Sequence Ontology (released on 11_22_2011) [21].

In addition to the two major alterations that resulted in the conversion of the model from one to three modules and in the incorporation of the knowledge from the controlled vocabularies, the model suffered several rounds of adjustments.

3 HCM Semantic Model

The resultant HCM model is composed by three modules:

- *Clinical Evaluation* - containing administrative concepts and clinical data elements that play a role in the diagnosis and the prognosis of HCM patients.
- *Genotype Analysis* - containing concepts associated with the genetic testing of biological samples.
- *Medical Classifications* - an auxiliary module containing medical standards used in the characterization of clinical elements such as patient symptoms.

[1] http://www.geneontology.org/formats/oboInOwl#hasDbXref

Table 1. Composition of the *Clinical Evaluation*, *Genotype Analysis*, and *Medical Classifications* modules in terms of: number of top-level concepts, total number of concepts, and number of data and object properties

Module	Top-level concepts	Total concepts	Properties
Clinical Evaluation	5	63	60
Genotype Analysis	7	19	39
Medical Classifications	2	4	2

Table 1 shows the composition of the three modules, both in number of concepts and properties. *Clinical Evaluation* is the largest, with a total of 63 concepts and approximately 60 object and data properties (Figures 2 and 3). *Genotype Analysis* contains 19 concepts and approximately 39 properties (Figure 4). Finally, *Medical Classifications* contains two high-level concepts (*Angina Classification* and *Heart Failure Classification*), each with one sub-concept, and a total of ten instances. As an example of this last module, Figure 5 shows the concept *Heart Failure Classification* and the data properties for one of its instances, *NYHA_Class2*.

The *Clinical Evaluation* (ce:) module imports the other two, *Genotype Analysis* (ga:) and *Medical Classifications* (mc:). The bridge between modules is made through the following non-hierarchical relationships (here represented as triples, where the central elements are object properties):

- ce:*Patient* ce:*hasBiologicalSample* ga:*Biological Sample*
- ce:*Biomarker Analysis* ce:*performedInBiologicalSample* ga:*Biological Sample*
- ce:*Angina* ce:*hasAnginaClassification* mc:*Angina Classification*
- ce:*Congestive Heart Failure* ce:*hasHeartFailureClassification* mc:*Heart Failure Classification.*

Patients' mutations can be identified through this relationship between *Clinical Evaluation* and *Genotype Analysis* since in the latter module a *Biological Sample* is connected with the mutations identified therein.

In terms of mappings to controlled vocabularies, SNOMED CT was used in the *Clinical Evaluation* module, the NCIt in the *Clinical Evaluation* and *Genotype Analysis* modules, the Gene Regulation Ontology and the Sequence Ontology in the *Genotype Analysis* module (see Table 2). More precisely, each vocabulary was considered in the following top-level concepts:

- SNOMED CT: *Clinical Finding* and *Observable Entity*
- NCIt: *Health Care Site*, *Person* and *Procedure* (from *Clinical Evaluation*); *Biological Sample*, *Gene*, *Mutation* and *Protein* (from *Genotype Analysis*)
- Gene Regulation Ontology: *Nucleic Acid Molecule*
- Sequence Ontology: *Primer*

Although the *Medical Classifications* module does not contain mappings to controlled vocabularies, its concepts are nonetheless linked to Web pages where their definition can be found.

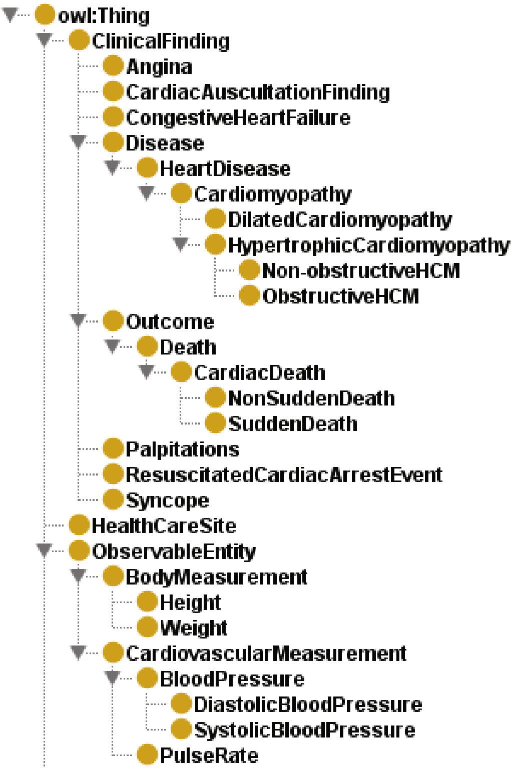
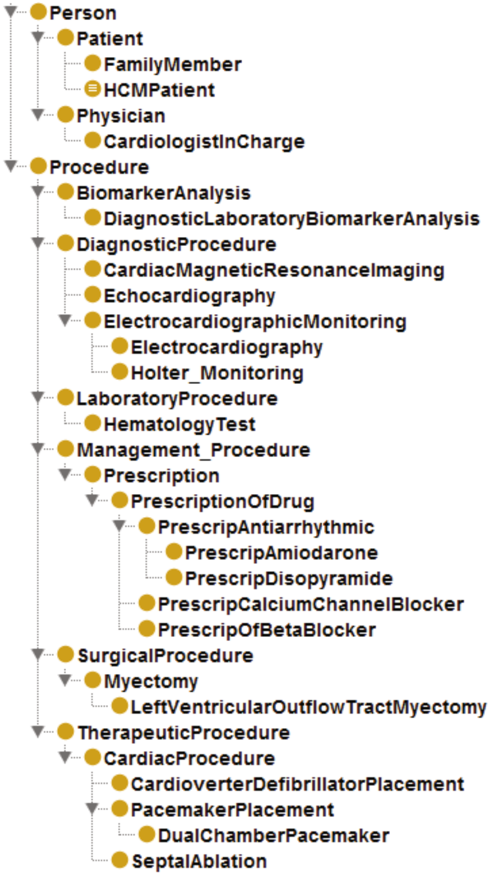


Fig. 2. Hierarchical structure of the *Clinical Evaluation* module, showing three of the top-level concepts (*Clinical Finding*, *Health Care Site* and *Observable Entity*) with all sub-concepts visible

4 Discussion

The decision to divide the model in modules was motivated by the observation that we wanted to represent two conceptually different types of knowledge: the knowledge related to patients, such as symptoms and treatments (represented in the *Clinical Evaluation* module); and the knowledge related to the analysis of biological samples collected from patients, such as amplification fragments and mutations (represented in the *Genotype Analysis* module). The third module, *Medical Classifications*, was added to represent standard medical classifications of any type, since these are independent from both patients and sample analysis. In terms of the use of the model in the final prognosis framework, this also means that we can easily provide two different views of the data: one centered on the patient, which is of interest for the medical doctors; and one centered on the biological samples, which is of interest for the molecular biologists.

Additionally, the modular development of the HCM model also facilitates its extension and reutilization. While the *Clinical Evaluation* module is the most



```
Person
  Patient
    FamilyMember
    HCMPatient
  Physician
    CardiologistInCharge
Procedure
  BiomarkerAnalysis
    DiagnosticLaboratoryBiomarkerAnalysis
  DiagnosticProcedure
    CardiacMagneticResonanceImaging
    Echocardiography
    ElectrocardiographicMonitoring
      Electrocardiography
      Holter_Monitoring
  LaboratoryProcedure
    HematologyTest
  Management_Procedure
    Prescription
      PrescriptionOfDrug
        PrescripAntiarrhythmic
          PrescripAmiodarone
          PrescripDisopyramide
        PrescripCalciumChannelBlocker
        PrescripOfBetaBlocker
  SurgicalProcedure
    Myectomy
      LeftVentricularOutflowTractMyectomy
  TherapeuticProcedure
    CardiacProcedure
      CardioverterDefibrillatorPlacement
      PacemakerPlacement
        DualChamberPacemaker
      SeptalAblation
```

Fig. 3. Hierarchical structure of the *Clinical Evaluation* module, showing two of the top-level concepts (*Person* and *Procedure*) with all sub-concepts visible. *Defined* classes, i.e. containing necessary and sufficient conditions, are indicated by a symbol with three horizontal lines.

specific and is best suited for the characterization of heart diseases, *Genotype Analysis* can be used in the context of any disease. In the case of *Medical Classifications*, although presently containing only two classes representing the classifications used by the medical experts, it can be expanded to include any standard or set of guidelines that refer to the medical aspects of HCM characterization or any other disease.

The use of controlled vocabularies proved to be advantageous on several levels: it saved us the work of creating a completely new model; it assisted us in identifying additional concepts and relations of interest; and it will facilitate the future addition of concepts since they can be searched in the vocabularies and easily integrated in their hierarchy. Nonetheless, the process was far from trivial.

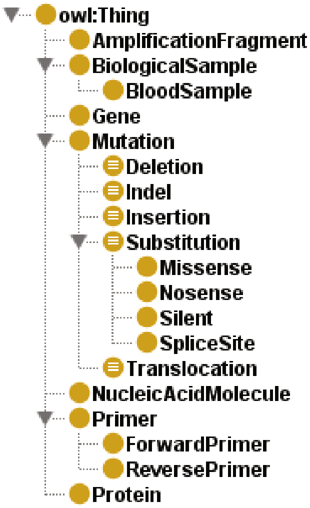
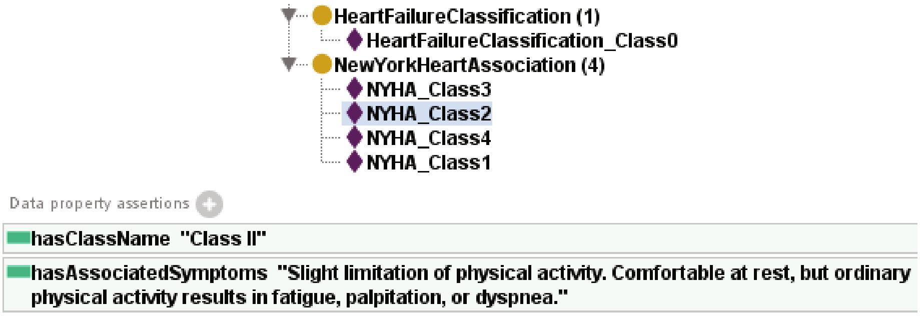


Fig. 4. Hierarchical structure of the *Genotype Analysis* module, showing all seven top-level concepts with all sub-concepts visible. *Defined* classes, i.e. containing necessary and sufficient conditions, are indicated by a symbol with three horizontal lines.

Table 2. Percentage of concepts from the HCM semantic model mapped to the following controlled vocabularies: SNOMED - Clinical Terms, NCI Thesaurus, Sequence Ontology and Gene Regulation Ontology. The percentages are indicated for the modules *Clinical Evaluation* and *Genotype Analysis*.

Module	Vocabulary (%)				Total (%)
	SNOMED CT	NCIt	SO	GRO	
Clinical Evaluation	42.9	42.9	-	-	85.8
Genotype Analysis	-	63.2	26.3	5.3	94.8

First of all, for the identification of the vocabularies we searched for concepts of interest on all the vocabularies available from BioPortal. This was a challenging task, since several vocabularies exist that fulfilled the requirement. After evaluating the most promising options, the initial list was progressively narrowed down until only those indicated remained. When this process was first started, we were not aware of the existence of the Biomedical Ontology Recommender service [22] available from BioPortal. However, we tested it afterwards and concluded that the vocabularies chosen coincided with the recommendations provided by the service. Additionally, the use of this service would have expedited considerably our work since it provides recommendations based on several concepts at the same time, and our searches were executed for one concept at a time.


HeartFailureClassification (1)
HeartFailureClassification_Class0
NewYorkHeartAssociation (4)
NYHA_Class3
NYHA_Class2
NYHA_Class4
NYHA_Class1
Data property assertions

hasClassName "Class II"

hasAssociatedSymptoms "Slight limitation of physical activity. Comfortable at rest, but ordinary physical activity results in fatigue, palpitation, or dyspnea."

Fig. 5. Representation of the concept *Heart Failure Classification* (from the *Medical Classifications* module) with: one instance; and the sub-concept *New York Heart Association* with its own four instances and the data properties of the instance *NYHA_Class2*

Secondly, several issues came to light during the development of the model related to:

- **Absent Concepts:** Inexistence of a concept of interest in the vocabulary.
- **Complexity:** Excess of concepts and of level of detail in general.
- **Placement:** Different possibilities concerning the placement of a concept in the hierarchy of the model.
- **Overlapping Regions:** Existence of overlapping concepts/regions of interest on different vocabularies.
- **Absent Textual Definitions:** Inexistence of textual definitions for concepts of interest.

The **Absent concepts** issue occurred both in the *Clinical Evaluation* and the *Genotype Analysis* modules. In the former module, we needed a concept *Cardiologist in charge* to represent the cardiologist that is primarily responsible for the HCM patient. According to the specifications of the biomedical experts guiding the development of the model, this cardiologist is the only medical doctor associated with the patient for this disease, and is responsible for every data element and evaluation represented in the model. Neither SNOMED CT nor NCIt provide such a representation, and the notions of "Physician" and of specific medical specialties such as "Cardiologist" are represented under *Occupation*, which can be interpreted as a label rather than a representation of a person. In this situation, we opted to use the concept *Person* from NCIt to aggregate *Patient* and *Physician*, and added *Cardiologist in Charge* as a sub-concept of *Physician*. In the *Genotype Analysis* module we needed to represent the *Translocation* and *Indel* sub-concepts of *Mutation*, as shown in Figure 4. While *Mutation* was mapped to the NCIt, this vocabulary does not include the indicated sub-concepts, and thus we mapped them to the Sequence Ontology.

The solution followed to deal with the **Complexity** of the controlled vocabularies, both in the form of number of concepts and detail of representation, was to consider only the concepts necessary for the description of the disease and for the structure of the model. The structure is particularly important for the mapping of the HCM model to external resources and for the future addition of concepts. An example of the complexity issue occurred with the concept *Procedure*. This concept is mapped to *Intervention or Procedure* from the NCIt, which contains thirteen sub-concepts, but we were interested in only five of them. If all thirteen were considered, the level of complexity of the model would be increased without any benefit for the end-users.

The **Placement** issue derived from our decision of not representing more than one parent per concept (i.e. multiparenting), even at the expense of a possible loss of detail. This decision was motivated by our intention of creating a model that would provide a straightforward experience to the biomedical experts when inputing or retrieving data (during the utilization of the prognosis framework), and thus avoid possible uncertainties due to multiple options. As such, we were occasionally forced to evaluate different possibilities for the placement of a concept in the hierarchy of the model. This occurred with concepts from SNOMED CT, in which situations we recurred to the NCIt to help us identify a solution common to both vocabularies. One such case occurred with *Syncope*, a *Clinical Finding* that is represented in SNOMED CT as a sub-concept of three different concepts: *Clinical history and observation finding*, *Finding by site* and *Disease*. In the HCM model we consider the concept *Clinical Finding* and its sub-concept *Disease*, and the decision was whether to place *Syncope* directly under the first-level *Clinical Finding* or the second-level *Disease*. In NCIt the concept is represented directly under the concept *Finding* and not under its sibling *Disease or disorder*, and consequently we chose to place it under *Clinical Finding* in the HCM model. Similar decisions were made for the concepts *Angina* and *Congestive heart failure*, which are sub-concepts of *Finding by site* and *Disease* in SNOMED CT, and of *Finding* in NCIt.

The **Overlapping regions** issue results from the existence of more than one vocabulary describing the same domain of knowledge. According to the accepted OBO Foundry [23] principle named "clearly delineated content" (FP005[2]), ontologies should be orthogonal to each other in order to enable the utilization of two different ontologies to define complementary perspectives on the same entities. In essence, we agree with this principle since the existence of a single ontology for a given domain would mean that anyone wanting to reuse it in an application semantic model would just have to follow it and consider the necessary knowledge. On the other hand, in light of our experience with the development of the HCM model, we consider that the availability of more than one vocabulary can be positive when no vocabulary is accepted as the single reference by the community.

An example of the overlapping regions in the *Clinical Evaluation* module occurred with the concept *Outcome*, a *Clinical Finding* with possible examples of

[2] http://www.obofoundry.org/wiki/index.php/FP_005_delineated_content

outcomes being decreased pain and death. *Clinical Finding* and its sub-concepts are mapped to SNOMED CT, but this vocabulary represents *Death* in a high-level class *Event*, which is not necessary for the HCM model. Moreover, NCIt has a concept *Outcome* under *Finding*, which also includes several sub-concepts relevant for the HCM model: *Death*, *Cardiac death*, *Sudden cardiac death* and *Non sudden cardiac death.* In this situation the decision was to consider *Outcome* and its sub-concepts from NCIt in the *Clinical Finding* concept, which is otherwise mapped to SNOMED CT.

Two other examples of the overlapping regions issue in the *Genotype Analysis* module involved the concepts *Primer* and *Nucleic acid molecule. Primer* is represented in the NCIt under *Drug, Food, Chemical or Biomedical Material* and without sub-classes. However, in the Sequence Ontology, a *Primer* is a *Sequence feature* with the two sub-classes *Forward Primer* and *Reverse Primer*, which were included in the HCM model. In the second situation, the concept *Nucleic Acid* was intended to represent actual nucleic acid molecules extracted from biological samples. While both the NCIt and the Sequence Ontology include the concept *Nucleic Acid*, neither define it suitably for our purposes: the former defines *Nucleic acids* as “A family of macromolecules”, whereas the latter defines *Nucleic acid* as “An attribute describing a sequence consisting of nucleobases bound to repeating units”. Consequently, we opted to use the Gene Regulation Ontology exclusively for its concept *Nucleic acid molecule*, witch is more suitably defined as a “A complex, high-molecular-weight biochemical macromolecule composed of nucleotide chains that convey genetic information”.

The overlapping regions issue is of particular importance given that using a domain representation that is unfamiliar to the end-users of the HCM prognosis framework may hinder significantly their acceptance of the framework.

The **Absent textual definitions** issue was perceived as a significant burden to the reuse of the affected concepts, since there where situations in which their intended use was not readily understandable. This was a common problem when using SNOMED CT, as this vocabulary lacks definitions for most of its concepts. For example when representing the concept *Cardiologist in Charge*, it was only possible to interpret the intended use of the concept *Cardiologist* based on the hierarchical organization of the vocabulary. By contrast, the NCIt has available detailed descriptions for the majority of its concepts, which provides a greater assistance when more complex decisions have to be made. This issue is not new, and has already been the subject of an OBO Foundry principle (FP 006 textual definitions[3]).

An issue particular to the development of the *Genotype Analysis* module occurred with the concepts *Nucleic acid molecule*, *Gene* and *Protein.* As represented in the Gene Regulation Ontology, these concepts are related with each other: *Gene* is represented under *DNA*, which in turn is a *Nucleic acid*; and *Nucleic acid* and *Protein* are both *Information biopolymer*(s) (“macromolecules that harbor biological information in their structures”). However, these relationships could not be conveyed in the HCM because what we want to represent

[3] http://www.obofoundry.org/wiki/index.php/FP_006_textual_definitions

under each concept is conceptually different: *Nucleic Acid Molecule*, the physical molecules; *Gene*, the list of genes associated with HCM (not the physical genes); and *Protein*, the list of proteins encoded by the genes associated with HCM (not the physical proteins).

5 Conclusions

Hypertrophic cardiomyopathy (HCM) is a complex genetic disease both in terms of diagnosis and prognosis, due to a great variability in terms of clinical manifestations and associated mutations. Furthermore, the presence of the same mutation in different individuals can result in very different clinical manifestations. Consequently, this disease is a good candidate for a translational medicine approach.

In this article we present a semantic data model that is the core element of a component of data representation and integration in our proposed prognosis framework for HCM. The data integrated with this component will be explored with data mining techniques to identify associations between clinical and genetic data. Our aim is that these associations might be used as guidelines to assist cardiologists in the prediction of the outcome of the disease for individual patients, in addition to existing guidelines [24]. In particular, we are interested in predicting the occurrence of sudden cardiac death.

The first step in the development of the model was the identification of the clinical and genetic data elements considered in the actual assessment of patients, in accordance with the practice of the medical and molecular biology experts with whom we collaborate.

The model was developed in OWL following a modular approach to facilitate its extension and reutilization. The concepts in all of the three modules that compose it (*Clinical Evaluation*, *Genotype Analysis* and *Medical Classifications*) are also mapped to external controlled vocabularies to facilitate the interaction with other systems. The current version of the semantic model includes mappings to the following four vocabularies: SNOMED CT, NCI Thesaurus, the Gene Regulation Ontology, and the Sequence Ontology. The use of these vocabularies was advantageous at various levels, but was not challenge-free. The solutions found resulted in a model that contains: mappings to more than one vocabulary; new concepts, not previously represented in any of the vocabularies; a minimum set of concepts necessary to describe the disease and to map it to external vocabularies; a hierarchical organization that results from more than one vocabulary. The solutions devised resulted from a compromise between the representations provided by the vocabularies and the vision of the domain conveyed by the biomedical experts that assisted in the development of the semantic model.

The model has been continuously evaluated in terms of correct representation of the domain of knowledge (performed by the biomedical experts) and in terms of consistency, with checks performed periodically normally after important alterations.

The semantic model is currently being populated with data from six Portuguese institutions: the *Hospitais da Universidade de Coimbra* (Coimbra), the *Centro de Cardiologia da Universidade de Lisboa*, the *Hospital da Luz* and the *Hospital de Sta. Cruz* (all three in Lisbon) provide the clinical data; the *Centro de Química Estrutural* of the *Instituto Superior Técnico of the Universidade Técnica de Lisboa* and the *Universidade Lusófona de Humanidades e Tecnologias* provide the genetic data. Future work includes the assessment of the effectiveness of the model to deal with real data.

Acknowledgments. This work was supported by the FCT through the Multiannual Funding Program, the doctoral grant SFRH/BD/65257/2009, the postdoctoral grant SFRH/BPD/20996/2004 and the SOMER project (PTDC/EIA-EIA/119119/2010).

References

1. Maron, B.J., Maron, M.S., Wigle, E.D., Braunwald, E.: The 50-Year History, Controversy, and Clinical Implications of Left Ventricular Outflow Tract Obstruction in Hypertrophic Cardiomyopathy: from Idiopathic Hypertrophic Subaortic Stenosis to Hypertrophic Cardiomyopathy. J. Am. Coll. Cardiol. 54, 191–200 (2009)
2. Alcalai, R., Seidman, J.G., Seidman, C.E.: Genetic Basis of Hypertrophic Cardiomyopathy: from Bench to the Clinics. J. Cardiovasc. Electrophysiol. 19, 104–110 (2008)
3. Harvard Sarcomere Mutation Database, http://genepath.med.harvard.edu/~seidman/cg3/
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: Tool for the Unification of Biology. Nat. Genet. 25, 25–29 (2000)
5. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED), http://www.ihtsdo.org/snomed-ct/
6. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Sci. Am., 29–37 (2001)
7. World Wide Web Consortium, http://www.w3.org/
8. RDF Primer, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
9. OWL Web Ontology Language Current Status, http://www.w3.org/standards/techs/owl#w3call
10. Gudivada, R.C., Qu, X.A., Chen, J., Jegga, A.G., Neumann, E.K., Aronow, B.J.: Identifying Disease-Causal Genes Using Semantic Web-based Representation of Integrated Genomic and Phenomic Knowledge. J. Biomed. Inform. 41, 717–729 (2008)
11. Agorastos, T., Koutkias, V., Falelakis, M., Lekka, I., Mikos, T., Delopoulos, A., Mitkas, P.A., Tantsis, A., Weyers, S., Coorevits, P., Kaufmann, A.M., Kurzeja, R., Maglaveras, N.: Semantic Integration of Cervical Cancer Data Repositories to Facilitate Multicenter Association Studies: the ASSIST Approach. Cancer Inform. 8, 31–44 (2009)

12. Colombo, G., Merico, D., Boncoraglio, G., Paoli, F.D., Ellul, J., Frisoni, G., Nagy, Z., van der Lugt, A., Vassányi, I., Antoniotti, M.: An Ontological Modeling Approach to Cerebrovascular Disease Studies: the NEUROWEB Case. J. Biomed. Inform. 43, 469–484 (2010)
13. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Technical report number KSL-01-05, Knowledge Systems, AI Laboratory, Stanford University (2001)
14. Protégé Ontology Editor, http://protege.stanford.edu
15. Hermit OWL Reasoner, http://hermit-reasoner.com/
16. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A.: BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. Nucleic Acids Res. 37, W170–W173 (2009)
17. Machado, C.M., Couto, F., Fernandes, A.R., Santos, S., Cardim, N., Freitas, A.T.: Semantic Characterization of Hypertrophic Cardiomyopathy Diseases. In: First Workshop on Knowledge Engineering, Discovery and Dissemination in Health, KEDDH 2010 (2010)
18. Sioutos, N., Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.L., Wright, L.W.: NCI Thesaurus: a Semantic Model Integrating Cancer-Related Clinical and Molecular Information. J. Biomed. Inform. 40, 30–43 (2007)
19. The Ontology of Clinical Research (OCRe), http://rctbank.ucsf.edu/home/ocre.html
20. Beisswanger, E., Lee, V., Kim, J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U.: Gene Regulation Ontology (GRO): Design Principles and Use Cases. St. Heal. T. 136, 9–14 (2008)
21. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M.: The Sequence Ontology: a Tool for the Unification of Genome Annotations. Genome Biol. 6, R44 (2005)
22. Jonquet, C., Musen, M.A., Shah, N.H.: Building a Biomedical Ontology Recommender Web Service. J. Biomed. Semantics 1(suppl. 1), S1 (2010)
23. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nat. Biotech. 25, 1251–1255 (2007)
24. Gersh, B.J., Maron, B.J., Bonow, R.O., Dearani, J.A., Fifer, M.A., Link, M.S., Naidu, S.S., Nishimura, R.A., Ommen, S.R., Rakowski, H., Seidman, C.E., Towbin, J.A., Udelson, J.E., Yancy, C.W.: ACCF/AHA Guideline for the Diagnosis and Treatment of Hypertrophic Cardiomyopathy: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 124, e783–e831 (2011)

Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation

Andreas Holzinger[1], Reinhold Scherer[2],
Martin Seeber[2], Johanna Wagner[2], and Gernot Müller-Putz[2]

[1] Institute for Medical Informatics, Statistics & Documentation, Research Unit HCI4MED
Medical University Graz, A-8036 Graz, Austria
andreas.holzinger@medunigraz.at
[2] Institute for Knowledge Discovery, Laboratory of Brain-Computer Interfaces
Graz University of Technology, A-8010 Graz, Austria

Abstract. Strokes are often associated with persistent impairment of a lower limb. Functional brain mapping is a set of techniques from neuroscience for mapping biological quantities (computational maps) into spatial representations of the human brain as functional cortical tomography, generating massive data. Our goal is to understand cortical reorganization after a stroke and to develop models for optimizing rehabilitation with non-invasive electroencephalography. The challenge is to obtain insight into brain functioning, in order to develop predictive computational models to increase patient outcome. There are many EEG features that still need to be explored with respect to cortical reorganization. In the present work we use independent component analysis, and data visualization mapping as tools for sensemaking. Our results show activity patterns over the sensorimotor cortex, involved in the execution and association of movements; our results further supports the usefulness of inverse mapping methods and generative models for functional brain mapping in the context of non-invasive monitoring of brain activity.

Keywords: Knowledge discovery, data mining, human-computer interaction, gait analysis, biomedical informatics, infomax independent component analysis.

1 Introduction, Methods and Experiment

Strokes are one of the most devastating of all neurological diseases, often leading to death or at least to physical impairment. There is a growing awareness of the potential for computer-mediated neuro-rehabilitation, which has led to various novel concepts for delivering these therapies (Harwin, Murgia & Stokes, 2011); advances in robotics along with an increased understanding of the latent neurologic potential for stroke recovery led to increasing use of robotic rehabilitation devices, having great potential to deliver efficient and reproducible therapies (Lo et al., 2010). Moreover, robotic therapy can be used to save time and energy for the therapist, making rehabilitation sessions more efficient, and rehabilitation protocols can be tailored to individual

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 166–168, 2012.

patients (Scherer et al., 2009). Rehab in practice works, therefore we hypothesize that there is a causal relationship between therapy and brain, supported by incorporation of functional clinical scores (Simonic et al., 2011). However, the biomedical experts are confronted with increased masses of highly complex, multi-variate and often weakly structured data (Holzinger, 2011). The integration of statistical methods and intelligent information visualization, to support sensemaking, thereby decision making is essential (Wong, Xu & Holzinger, 2011).

In this work we applied independent component analysis (ICA), for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals (Comon, 1994), (Boehm, Faloutsos & Plant, 2008). The goal is to express a set of random variables as linear combinations of statistically independent component variables. In our experiment the electro-cortical activity was recorded from 120 channels, equally distributed over the participant's head during walking in a robotic gait orthosis (Riener et al., 2010). Current neuroimaging methods make it very difficult to record bio-signals during whole body movements; as fMRI, MEG and EEG are prone to artefact contamination. Our EEG data was decomposed into 120 independent components and EEG data analysis was performed in Matlab using EEGLAB 8.0.3.5b (Delorme & Makeig, 2004). An equivalent current dipole was calculated for each independent component scalp projection using a standardized three-shell boundary element head model (BEM) (Oostenveld & Oostendorp, 2002). We expected to find a source in central midline areas related to foot movements during walking. This was achieved by the application of expert knowledge about the neurophysiology and the representation of brain processes in the EEG. The independent components were visually inspected by the domain expert and classified into brain related sources and components representing artefacts.

2 Results, Conclusion and Future Research

Our results showed patterns over the sensorimotor area that is involved in the execution and association of movements. To acquire these results we used big data sets of various sources including EEG-data (electrical potentials), MRI Images and 3D-space coordinates of the electrodes positions. Consequently, we acquired a visual representation of the brain activity and mapped this on the subjects' individual anatomy. Infomax independent component analysis proved again to be powerful. Image segmentation, forward- and inverse modeling were necessary to get a set knowledge. Since each of these tools need a set of user defined parameters or boundary conditions the expertise of the user is crucial to provide neurophysiological meaningful results. Constituent and essential was the interactive involvement of an domain expert, who visually inspected and classified brain related sources and components representing artefacts. We consider our work as a small, but important step towards enhancing stroke rehabilitation – to get from data to patient centered therapy. Our results confirm the usefulness of inverse mapping methods for functional brain mapping and generative models in the context of non-invasive monitoring of brain activity. In other terms we showed that we can create knowledge out of big data, by combination of several computational tools and human expertise that is meaningful for known patterns. The next goal of our future research is to interpret this generated knowledge for unknown patterns to provide more information in the context of modeling of the plasticity of the brain after injury, e.g. stroke.

Acknowledgement. This work was funded by Land Steiermark project bci4rehab.

References

Harwin, W., Murgia, A., Stokes, E.: Assessing the effectiveness of robot facilitated neurorehabilitation for relearning motor skills following a stroke. Medical and Biological Engineering and Computing 49(10), 1093–1102 (2011)

Lo, A.C., Guarino, P.D., Richards, L.G., Haselkorn, J.K., Wittenberg, G.F., Federman, D.G., Ringer, R.J., Wagner, T.H., Krebs, H.I., Volpe, B.T., Bever, C.T., Bravata, D.M., Duncan, P.W., Corn, B.H., Maffucci, A.D., Nadeau, S.E., Conroy, S.S., Powell, J.M., Huang, G.D., Peduzzi, P.: Robot-Assisted Therapy for Long-Term Upper-Limb Impairment after Stroke. New England Journal of Medicine 362(19), 1772–1783 (2010)

Scherer, R., Pradhan, S., Dellon, B., Kim, D., Klatzky, R., Matsuoka, Y.: Characterization of multi-finger twist motion toward robotic rehabilitation. In: ICORR 2009, Kyoto (Japan), pp. 812–817. IEEE (2009)

Simonic, K.M., Holzinger, A., Bloice, M., Hermann, J.: Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. In: Proceedings of Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, pp. 550–554. IEEE (2011)

Holzinger, A.: Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction. In: Baghaei, N., Baxter, G., Dow, L., Kimani, S. (eds.) Proceedings of INTERACT 2011 Workshop: Promoting and Supporting Healthy Living by Design. IFIP, Lisbon (Portugal), pp. 5–7 (2011)

Wong, B.L.W., Xu, K., Holzinger, A.: Interactive Visualization for Information Analysis in Medical Diagnosis. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 109–120. Springer, Heidelberg (2011)

Comon, P.: Independent Component Analysis, a new concept? Signal Processing 36(3), 287–314 (1994)

Boehm, C., Faloutsos, C., Plant, C.: Outlier-robust clustering using independent components. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, Canada, pp. 185–198. ACM (2008)

Riener, R., Lünenburger, L., Maier, I.C., Colombo, G., Dietz, V.: Locomotor Training in Subjects with Sensori-Motor Deficits: An Overview of the Robotic Gait Orthosis Lokomat. Journal of Healthcare Engineering 1(2), 197–216 (2010)

Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134, 9–21 (2004)

Oostenveld, R., Oostendorp, T.F.: Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. Human Brain Mapping 17, 179–192 (2002)

The Database of the Cardiovascular System Related Signals

Jan Havlík[1], Lucie Kučerová[1], Imrich Kohút[1],
Jan Dvořák[1], and Vratislav Fabián[2]

[1] Department of Circuit Theory, Faculty of Electrical Engineering
Czech Technical University in Prague, Technická 2, CZ-16627 Prague 6
xhavlikj@fel.cvut.cz
[2] Department of Physics, Faculty of Electrical Engineering
Czech Technical University in Prague, Technická 2, CZ-16627 Prague 6

Abstract. The paper presents the design and development of a signal database used for the determination of hemodynamic parameters. The signals from the database are used for the evaluation of new algorithms determining hemodynamic parameters. The signal database consists of synchronously obtained independent signals – the records of oscillometric pulsations, the records of electrocardiograms (ECG) and the records of photoplethysmogram (PPG). Currently the signal database consists of signals from about 65 persons. The signals were recorded from persons in wide age range from 19 to 94 years. The signals were stored anonymously, but each set of signals was labeled by the ID number and the anamnestic data.

Keywords: hemodynamic parameters, oscillometric signals, electrocardiography, photoplethysmography.

1 Introduction

The paper presents the design and development of a signal database used for the determination of hemodynamic parameters. The database is designed as a set of signals synchronously obtained using our own medical device. The signals from the database are used for the evaluation of new algorithms determining hemodynamic parameters. The aim of the research is to design new methods for primary screening of atherosclerosis based on the strictly non-invasive methods and without need to use advanced imaging methods.

2 Signal Database

The signal database consists of synchronously obtained independent signals – the records of oscillometric pulsations obtained with the arm cuff (the pulsations were obtained both during inflation and deflation of the cuff), the records of electrocardiograms (ECG) obtained by three leads system (three electrodes

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 169–170, 2012.

for measuring signals from Einthoven triangle and the electrode placed on the right leg for the noise reduction) and the records of photoplethysmogram (PPG) obtained using the plethysmography sensor placed on the index finger.

Currently the signal database consists of signals from about 65 persons. The signals were recorded from persons in wide age range from 19 to 94 years, typically from the students of the university and from elderly persons in one of the Prague's senior houses. The signals were stored anonymously in the database. Each set of signals was labeled by the ID number and the anamnestic questionnaire labeled with the same ID was filled in with the measured person. The probands were asked for the age, sex, weight, height, life style, prescribed medicaments, smoking and related anamnestic data (hypertension or hypotension, diabetes, cardiovascular illnesses, respiratory illnesses etc.). Before the measurement each person was instructed about the method of measurement and about the aim of the research and subsequently the informed consent was signed.

3 Conclusion

The complex signal database of oscillometric signals, ECG and PPG signals has been created. The database consists of unique clinical data which have not been collected previously. The database is mainly intended for the research in the field of hemodynamic parameters and the primary screening of atherosclerosis. The signals from the database will be used for the evaluation of the algorithms developed for prediction of cardiovascular diseases. Several hemodynamic parameters such as pulse wave velocity (PWV), arterial stiffness index (ASI) or cardio-ankle vascular index (CAVI) will be determined using the signals.

Acknowledgement. This work has been supported by the grant No. SGS11/153/OHK3/3T/13 of the Czech Technical University in Prague and by the research program No. MSM 6840770012 of the Czech Technical University in Prague (sponsored by the Ministry of Education, Youth and Sports of the Czech Republic).

References

1. Lopes, A.A., OLeary, P.W.: Measurement, interpretation and use of hemodynamic parameters. Cardiology in the Young 19(suppl. S1), 8 (2009), http://dx.doi.org/10.1017/S1047951109003886
2. National Heart Lung and Blood Institute: What is atherosclerosis?, http://www.nhlbi.nih.gov/health/health-topics/topics/atherosclerosis/
3. Tholl, U., Forstner, K., Anlauf, M.: Measuring blood pressure: pitfalls and recommendations. Nephrology Dialysis Transplantation 19(4), 766 (2004)

Patient Monitoring Using Bioimpedance Signal

Jan Havlík, Ondřej Fousek, and Miroslav Ložek

Department of Circuit Theory, Faculty of Electrical Engineering,
Czech Technical University in Prague, Technická 2, CZ-16627 Prague 6
xhavlikj@fel.cvut.cz

Abstract. The paper presents the development of prototype system for measurement of vital functions using the transthoracic bioimpedance. The measurement of the bioimpedance is a non-invasive method providing the information about the body composition, hearth rate, blood flow, breathing etc.

The design and realization of a simple four electrode system have been done. The device is designed as a combination of a signal generator with stabilized current output and a measuring amplifier based on the AD620 amplifier.

The output signal includes information about the basic vital signs and could be used as a part of telemonitoring system for the elderly and persons after the organs failure. The device will be used for research and educational purposes in the Smart Home facility at the Czech Technical University in Prague.

Keywords: bioimpedance, vital functions, cardiac output, blood flow.

1 Introduction

The paper presents the design and realization of prototype system for patient monitoring using transthoracic bioelectric impedance signal (frequently shortened as bioimpedance).

Bioimpedance is a response of a human body (generally of any living organism) to an externally driven electric current. Based on the Ohm law the current pass through the body evokes the decrease of the voltage on the body. From the known values of the current and the voltage the bioimpedance could be determined. The bioimpedance measurement is a non-invasive method that provides information about the body composition, hearth rate, blood flow, breathing etc. In medicine the bioimpedance signal is often used for the non-invasive measurement of the cardiac output (or minute volume) and for the determination of the blood flow (without Doppler sonography).

Based on the information above the bioimpedance measurement could substitute many other measurements, for example the electrocardiography (ECG) measurement of heart activity, the plethysmography (PPG) measurement and the measurement of cardiac output (CO) in one way. Regardless in the field of a patient monitoring, telemonitoring of vital signs, ambient assisted living and smart homes the bioimpedance is wrongfully omitted.

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 171–172, 2012.

2 Realization

The paper presents the simple four electrode system for the bioimpedance measurement fully capable to give the information about the basic vital functions. The device is designed as a simple combination of the driven signal generator and the measuring amplifier. The signal generator consists of the sinus generator and the voltage/current converter with an operational amplifier. It means the generator has the stabilized current output. The measuring amplifier is based on the AD620 amplifier from Analog Devices. The output of the device is used as an input signal for our own telemonitoring system, the signal is then preprocessed by the system and wirelessly transfered to the PC.

3 Results

The basic hemodynamic parameters like hearth rate and the information about the cardiac output is determined based on the measured signal. The outputs could be used as a part of vital signs telemonitoring system for elderly and persons during the convalescence after the organs failure (hearth failure, brain stroke). It will be used both for research and education in the Smart Home facility at the Czech Technical University in Prague.

Acknowledgement. This work has been supported by the grant No. F3a 2177/2012 of the University Development Foundation and by the research program No. MSM 6840770012 of the Czech Technical University in Prague (sponsored by the Ministry of Education, Youth and Sports of the Czech Republic).

References

1. Cybulski, G.: Ambulatory Impedance Cardiography: The Systems and their Applications. Lecture Notes in Electrical Engineering. Springer (2011)
2. Grimnes, S., Martinsen, O.G.: Bioimpedance and Bioelectricity Basics. Academic Press (2000)
3. Lababidi, Z., Ehmke, D.A., Durnin, R.E., Leaverton, P.E., Lauer, R.M.: The first derivative thoracic impedance cardiogram. Circulation 41(4), 651–658 (1970)
4. Nyboer, J.: Electrical impedance plethysmography; a physical and physiologic approach to peripheral vascular study. Circulation 2(6), 811–821 (1950)
5. Sakamoto, K., Muto, K., Kanai, H., Iizuka, M.: Problems of impedance cardiography. Medical and Biological Engineering and Computing 17(6), 697–709 (1979)
6. Stevanovic, P., Stepanovic, R., Radovanovic, D., Bajec, D., Perunovic, R., Stojanovic, D., Stevanovic, D.: Thoracic electrical bioimpedance theory and clinical possibilities in perioperative medicine. Journal for Intensive Care and Emergency Medicine, 22–27 (2008)

Author Index

Abelha, António 1
Augusto, Alexandre B. 54

Basagiannis, George 69
Boeva, Veselka 84
Bursa, Miroslav 31

Ceci, Michelangelo 109
Chudacek, Vaclav 31
Correia, Manuel E. 54
Couto, Francisco M. 151
Croset, Samuel 124
Cruz, Jorge 16
Cruz-Correia, Ricardo 54

Dvořák, Jan 169

Fabián, Vratislav 169
Fernandes, Alexandra R. 151
Ferreira, Ana 54
Folino, Francesco 99
Fousek, Ondřej 171
Freitas, Ana T. 151
Fumarola, Fabio 109

Girardi, Dominic 40
Giretzlehner, Michael 40
Guzzi, Pietro Hiram 109

Havlík, Jan 169, 171
Hoehndorf, Robert 124
Holzinger, Andreas 166
Hristoskova, Anna 84
Huser, Martin 31

Janku, Petr 31

Kohút, Imrich 169
Kozak, Karol 139

Kučerová, Lucie 169
Küng, Josef 40

Lhotska, Lenka 31
Ložek, Miroslav 171

Machado, Catia M. 151
Machado, José 1
Mandreoli, Federica 109
Martoglia, Riccardo 109
Masciari, Elio 109
Mecella, Massimo 109
Müller-Putz, Gernot 166

Paschou, Mersini 69
Penzo, Wilma 109
Pizzuti, Clara 99
Portela, Filipe 1

Riley, Luke 124
Rua, Fernando 1

Sakkopoulos, Evangelos 69
Santos, Manuel Filipe 1
Santos, Maribel Yasmina 16
Santos, Susana 151
Santos-Pereira, Cátia 54
Scherer, Reinhold 166
Seeber, Martin 166
Silva, Álvaro 1
Sourla, Efrosini 69
Spilka, Jiri 31

Teles de Araújo, Artur 16
Tsakalidis, Athanasios 69
Tsiporkova, Elena 84

Wagner, Johanna 166
Wyner, Adam 124