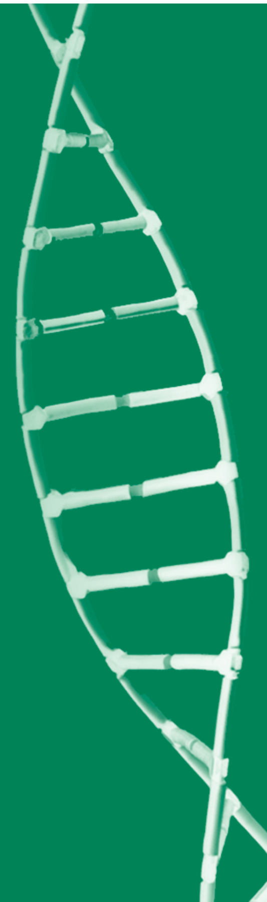
LNBI 7845

Leif E. Peterson
Francesco Masulli
Giuseppe Russo (Eds.)

Computational Intelligence Methods for Bioinformatics and Biostatistics

9th International Meeting, CIBB 2012
Houston TX, USA, July 2012
Revised Selected Papers



Springer

Lecture Notes in Bioinformatics 7845

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Leif E. Peterson   Francesco Masulli
Giuseppe Russo (Eds.)

Computational Intelligence Methods for Bioinformatics and Biostatistics

9th International Meeting, CIBB 2012
Houston, TX, USA, July 12-14, 2012
Revised Selected Papers

Springer

Volume Editors

Leif E. Peterson
Cornell University
Center for Biostatistics, TMHRI, Weill Cornell Medical College
6565 Fannin Street, Mary Gibbs Jones Hall, Houston, TX 77030, USA
E-mail: peterson.leif@ieee.org

Francesco Masulli
University of Genoa
DIBRIS
Via Dodecaneso 35, 16146 Genoa, Italy
E-mail: francesco.masulli@unige.it

Giuseppe Russo
Temple University
Center for Biotechnology
Sbarro Institute for Cancer Research and Molecular Medicine
1900 N 12th Street, BioLife Science Bldg., Philadelphia, PA 19122, USA
E-mail: russo@temple.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-38341-0 e-ISBN 978-3-642-38342-7
DOI 10.1007/978-3-642-38342-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013937846

CR Subject Classification (1998): I.5, I.2.3, I.2.6, H.2.8, I.4, F.1, J.3, F.2, G.1.6

LNCS Sublibrary: SL 8 – Bioinformatics

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

This volume contains a selection of the best contributions delivered at the 9^{th} International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2012) held at The Methodist Hospital Research Institute (TMHRI), Houston, Texas (USA), during 12–14 July, 2012.

The CIBB meeting series is organized by the Special Interest Group on Bioinformatics and Intelligence of the International Neural Network Society (INNS) to provide a forum open to researchers from different disciplines to present and discuss problems concerning computational techniques in bioinformatics, systems biology and medical informatics with a particular focus on neural networks, machine learning, fuzzy logic, and evolutionary computational methods. Until 2012, CIBB meetings were held annually in Italy with an increasing number of participants: From 2004 to 2007, CIBB had the format of a special session of larger conferences, namely, WIRN 2004 in Perugia, WILF 2005 in Crema, FLINS 2006 in Genoa, and WILF 2007 in Camogli. Given the great success of the special session at WILF 2007 that included 26 strongly rated papers, the Steering Committee decided to turn CIBB into an autonomous conference starting with the 2008 edition in Vietri. The following editions in Italian venues were held in Genoa (2009), Palermo (2010) and Gargnano (2011). CIBB 2012 was the first edition organized outside Italy, and attracted 23 paper submissions from national and international research groups. A rigorous peer-review selection process was applied to ultimately select the papers included in the program of the conference. This volume includes the best contributions presented at the conference.

The success of CIBB 2012 is to be credited to the contribution of many people. Firstly, we would like to thank the organizers of the special sessions for attracting so many strong papers, which extended the focus of the main topics of CIBB. Second, particular thanks are due to the Program Committee members and reviewers for providing high-quality reviews. We would like to thank the keynote speakers Jim Bezdek (University of West Florida, USA), Elia Biganzoli (University of Milan, Italy), and Doug Robinson (SAS-JMP Genomics, Cary, USA).

February 2013

Leif E. Peterson
Francesco Masulli
Giuseppe Russo

Organization

The 9th CIBB meeting was a joint operation of the Task Force on Neural Networks of the IEEE-CIS Technical Committee on Bioinformatics and Bioengineering, and of two INNS Special Interest Groups: the SIG on Bioinformatics and Intelligence and the SIG on Biopattern. This operation was in collaboration with the Italian Neural Networks Society, The Methodist Hospital Research Institute, Houston, Texas, USA, the Sbarro Health Research Organization, Philadelphia, Pennsylvania, USA, the Department of Computer and Information Sciences, University of Genoa, Italy, and the Department of Mathematics and Computer Science, University of Salerno, Italy.

Conference Chairs

Leif E. Peterson	The Methodist Hospital Research Institute, Houston, USA
Francesco Masulli	University of Genoa, Italy; Temple University, Philadelphia, USA
Giuseppe Russo	Temple University, Philadelphia, USA

CIBB Steering Committee

Pierre Baldi	University of California, Irvine, USA
Elia Biganzoli	University of Milan, Italy
Alexandru Floares	Oncological Institute Cluj-Napoca, Romania
Jonathan Garibaldi	University of Nottingham, UK
Nikola Kasabov	Auckland University of Technology, New Zealand
Francesco Masulli	University of Genova, Italy
Leif Peterson	TMHRI, Houston, Texas, USA
Roberto Tagliaferri	University of Salerno, Italy

Biostatistics Technical Chair

Federico Ambrogi	University of Milan, Italy

Bioinformatics Technical Chair

Vassilis Plagianakos	University of Central Greece, Greece

Program Committee

Claudia Angelini	Istituto per le Applicazioni del Calcolo IAC-CNR, Italy
Sansanee Auephanwiriyakul	Chiang Mai University, Thailand
Sanghamitra Bandyopadhyay	Indian Statistical Institute, Kolkata, India
Gilles Bernot	University of Nice-Sophia, Antipolis, France
Chengpeng Bi	Childrens Mercy Hospital, Kansas City, USA
Mario Cannataro	University of Magna Graecia, Italy
Virginio Cantoni	Università di Pavia, Italy
Xue-Wen Chen	University of Kansas, Lawrence, USA
Adele Cutler	Utah State University, Logan, USA
Giuseppe Di Fatta	University of Reading, UK
Paolo Decuzzi	TMHRI, Houston, USA
Joaquin Dopazo	Centro De Investigacion Prencipe Felipe, Valencia, Spain
Angelo Facchiano	Istituto di Scienze dell'Alimentazione - CNR, Avellino, Italy
Maurizio Filippone	University of Glascow, UK
Enrico Formenti	University of Nice-Sophia, Antipolis, France
Leonardo Franco	Universidad de Malaga, Spain
Christoph Friedrich	University of Applied Science and Arts, Dortmund, Germany
Saman Halgamuge	University of Melbourne, Australia
Emmanuel Ifeachor	University of Plymouth, UK
Paulo Lisboa	Liverpool John Moores University, UK
Elena Marchiori	Radboud University, Nijmegen, The Netherlands
Luciano Milanesi	Istituto di Tecnologie Biomediche, ITB-CNR, Milan, Italy
Taishin Nomura	Osaka University, Japan
Riccardo Rizzo	ICAR-CNR, Palermo, Italy
Paolo Romano	National Cancer Research Institute, Genoa, Italy
Stefano Rovetta	University of Genova, Italy
Jianhua Ruan	University of Texas, San Antonio, USA
Andrey Rzhetsky	University of Chicago, USA
Mika Sato-Ilic	University of Tsukuba, Japan
Jennifer Smith	Boise State University, USA
Giorgio Valentini	Università degli Studi di Milano, Italy
Alfredo Vellido	Universidad Politecnica de Cataluna, Spain
Yanqing Zhang	Georgia State University, Atlanta, USA

Additional Reviewers

Michael Vrahatis
Konstantinos Parsopoulos
Ilias Maglogiannis
Nicos Pavlidis
Michael Epitropakis
George Magoulas
Anastasios Bezerianos
Aristotelis Chatziioannou
Zoi Litou
Bessy Iconomidou
Stefanos Bonovas
Sotiris Bersimis
Sophia Kossida
Sotiris Tasoulis

Local Scientific Secretary

Leif E. Peterson — Center for Biostatistics, TMHRI, Houston, Texas

Financing Institutions

Center for Biostatistics, TMHRI, Houston, Texas, USA
SAS-JMP Genomics, Cary, North Carolina, USA

Table of Contents

I Relativistic Heavy Ions and DNA Damage

A Theoretical Analysis of Visual Distributions of Ionizing-Radiation-Induced Foci in Human Cells by Heavy Ions 1
Artem L. Ponomarev and Francis A. Cucinotta

Multiple CPU Computing: The Example of the Code RITRACKS 12
Ianik Plante and Francis A. Cucinotta

Modeling the Depressed Hematopoietic Cells for Immune System under Chronic Radiation 26
Shaowen Hu and Francis A. Cucinotta

II Image Segmentation

Feature-Based Medical Image Registration Using a Fuzzy Clustering Segmentation Approach 37
Hassan Mahmoud, Francesco Masulli, and Stefano Rovetta

A Novel Approach Based on Joint Optimization of Alignment and Statistical Surface Representation with Wavelet Transform for CBCT Segmentation 48
Yu-Bing Chang, Peng Yuan, Tai-Hong Kuo, Zixiang Xiong, Jaime Gateno, James J. Xia, and Xiaobo Zhou

III Proteomics

Searching Structural Blocks by SS Exhaustive Matching 57
Virginio Cantoni, Alessio Ferone, Ozlem Ozbudak, and Alfredo Petrosino

Characterizing Intermediate Conformations in Protein Conformational Space 70
Rosanne Vetro, Nurit Haspel, and Dan Simovici

A Method of Extracting Sentences Containing Protein Function Information from Articles by Iterative Learning with Feature Update ... 81
Kazunori Miyanishi and Takenao Ohkawa

IV RNA and DNA Sequence Analysis

Non-coding RNA Covariance Model Combination Using Mixed Primary-Secondary Structure Alignment 95
Jennifer A. Smith

A Study of Compression–Based Methods for the Analysis of Barcode Sequences 105
Massimo La Rosa, Antonino Fiannaca, Riccardo Rizzo, and Alfonso Urso

V RNA, DNA, and SNP Microarrays

Assessing Agreement between microRNA Microarray Platforms via Linear Measurement Error Models 117
Niccolò Bassani, Federico Ambrogi, and Elia Biganzoli

Extracting Key Pathways from Gene Signature and Genetic Aberrations in Subtypes of Cancer 132
Peikai Chen, Yubo Fan, Tsz-kwong Man, Ching C. Lau, Y.S. Hung, and Stephen T.-C. Wong

VI Semi-Supervised/Unsupervised Cluster Analysis

Investigating Distance Metrics in Semi-supervised Fuzzy c-Means for Breast Cancer Classification 147
Daphne Teck Ching Lai and Jonathan M. Garibaldi

Two Covariances Harnessing Fuzzy Clustering Based PCA for Discrimination of Microarray Data 158
Mika Sato-Ilic

Random Matrix Theory and Covariance Matrix Filtering for Cancer Gene Expression 173
Leif E. Peterson and Charles E. Ford

Author Index 185

A Theoretical Analysis of Visual Distributions of Ionizing-Radiation-Induced Foci in Human Cells by Heavy Ions

Artem L. Ponomarev[1,2] and Francis A. Cucinotta[2]

[1] USRA, 3600 Bay Area Blvd., Houston, TX 77058
artem.l.ponomarev@nasa.gov
[2] NASA Lyndon B. Johnson Space Center, 2101 NASA Parkway, Houston, TX 77058

Abstract. Our purpose is to improve counting of DNA damage foci for high-LET (linear energy transfer) irradiation using computer modeling**.** The analysis of patterns of DNA RIFs (radiation-induced foci) produced by high-LET Fe ions was conducted by a Monte Carlo model that superimposes the heavy ion track structure with the human genome on the level of chromosomes. The image segmentation algorithm is used to compare the Monte Carlo data with experimental images of RIFs. In the experiment, we conducted the enumeration of radiation-induced foci using immunofluorescence using proteins that detect DNA damage. The model predicts the spatial and genomic distributions of DNA DSBs (double strand breaks) in a cell nucleus for a particular dose of radiation that can be compared to the visible distributions of RIFs. We used the model to do analyses for three irradiation scenarios: 1) the ions were oriented perpendicular to the flattened nuclei in a cell culture monolayer; 2) the ions were parallel to that plane; and 3) round nucleus. For these scenarios the statistics and spatial distribution of regions of densely arranged foci, termed DNA foci chains, were predicted numerically using this model. We showed that DSB clustering needs to be taken into account to determine the true DNA damage foci yield, which helps to determine the DSB yield. Using the model analysis, a researcher can refine the DSB yield per nucleus per particle. We showed that purely geometric artifacts, present in the experimental images, can be analytically resolved with the model, and that the quantization of track hits and DSB yields can be provided to the experimentalists.

Keywords: NASARadiationTrackImage model, ionizing radiation, human cells, DNA damage foci, DNA double-strand breaks, image segmentation, Monte Carlo method.

1 Introduction

Heavy ion radiation will be encountered during future space missions [1] and in hadron radiation therapy [2]. The yield of DNA double-strand breaks (DSBs) from heavy ion radiation is an important measure of cellular damage. DSBs can lead to cell death, chromosomal aberrations, mutations, and genomic instability, which can

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 1–11, 2013.

contribute to late effects such as cancer. The DSB yield is sometimes estimated by counting radiation-induced foci (RIFs) [3,4,5]. One component of these foci is the phosphorylated histone variant H2AX (histone variant 2AX), denoted as gamma-H2AX, which rapidly accumulates at sites of DNA DSBs [3,6,7,8]. The DSB yield per cell at a given dose can be roughly determined by counting foci at a given time following radiation exposure. However, because foci formation is a biochemical kinetic process involving both formation and loss of foci [9] there is never an exact one-to-one correspondence between the number of DSBs and foci even for X-rays. Also, low level background foci may be present that are not associated with DSBs but are the result of other cellular processes such as DNA replication [10]. In addition, multiple DSBs may be visible as a single focus due to DNA supercoiling [11]. These DSBs appear to be closely associated physically, but are actually distally located in space along a DNA molecule. For heavy ions clustering of damage including DSBs along the trajectory of the ion leads to further complexities in relating foci counts to damage numbers.

This paper reviews our work that addresses our approaches to overcome the difficulties that arise in accurate enumeration of RIFs that result from merged and closely aligned DNA DSBs that often result from heavy ion radiation, where damaged areas appear as streaks of merged foci [12]. This blurring, or merging, of foci is practically absent in cells irradiated with X-rays or γ-ray -- both low-LET (linear energy transfer) radiations, as they are more homogenously distributed in experimental images. We have termed this particular artifact a foci chain, meaning an optically visible chain of several foci. A strict algorithmic definition will be given to the notion of a foci chain with the intent to quantitate these objects and to infer the "true" DSB yield. The algorithm simulates a three-dimensional (3D) cell nucleus and generates the foci in three dimensions. The application of the model to the maximum projection images, where the foci are recorded into a two-dimensional (2D) image, is given by the analysis of foci chains in a 2D plane corresponding to an experimental image. To study the relation of DNA DSBs to foci we will assume a simplified concept of a DSB: it is a DNA break within a DNA locus of approximately 2 kilo base pairs (kbp) in genomic content. The justification for this approximation is given in [11]. Multiple DSBs within the same locus, should they occur, will be counted as single events with this approximation. Below we show the key results that would enable an experimentalist to quantitate heavy-ion-induced foci accurately and help to resolve streaks of foci in images [12]. This work provides an approach to improve the count of foci and information on how to apply these findings to the determination of DSB yields in human cells. Another application of our visual analysis approaches developed herein was made in [13].

2 Materials and Methods

Cells were exposed to 1,000 MeV/n Ti ions at the NASA (National Aeronautics and Space Administration) Space Radiation Laboratory at Brookhaven National Laboratory (NSRL), with the beamline parallel to the monolayer of cells. At 30 minutes after exposure, cultures were fixed with 4% paraformaldehyde for 10 minutes at room temperature and were washed 2 times with phosphate buffered saline (PBS). Cultures were shipped back to JSC (NASA Lyndon B. Johnson Space Center) for further analysis.

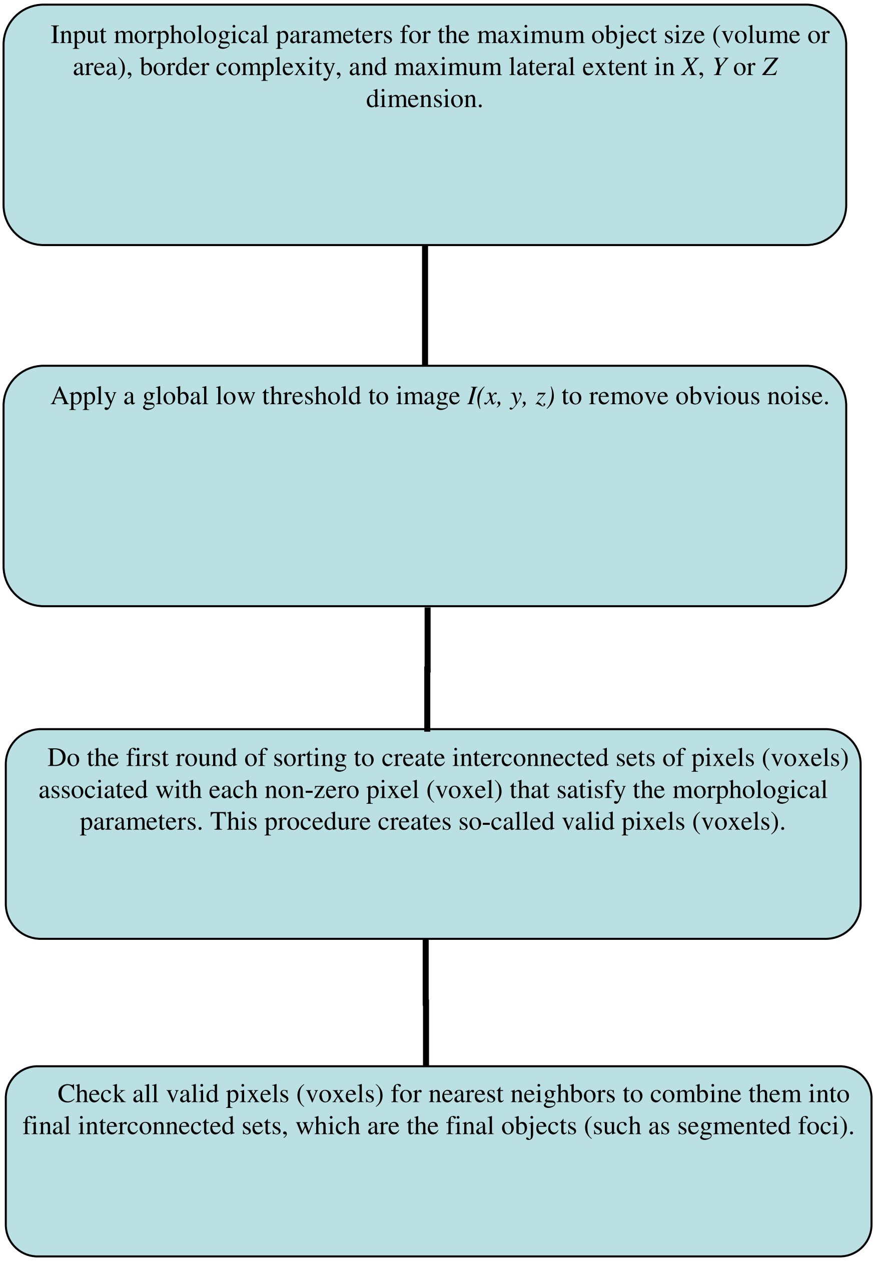
For immunostaining, cells were washed in PBS, permeablized using 0.1% Triton X-100 for 3 minutes, followed by washing 3 times in PBS. Cells were blocked with 1% bovine serum albumin /0.1% Nonidet P-40 for 30 minutes at room temperature, and then incubated for 1 hour at room temperature with mouse antibody against pSer1981 ataxia telangiectasia, mutated (ATM, or ataxia telangiectasia mutated) (Rockland Immunochemicals, Gilbertsville, Pennsylvania, USA) and rabbit antibody against γ-H2AX (Millipore, Billerica, Massachussetts, USA). Following three washes with PBS (phosphate buffered saline), cells were incubated with secondary antibodies Alexa488 anti-mouse and Alexa594 anti-rabbit (Molecular Probes/Invitrogen Carlsbad, California US) for 30 minutes at room temperature, washed 3 times with PBS and mounted using ProlongGold containing 4',6-diamidino-2-phenylindole (Invitrogen). Images were collected and recorded using a Leica TCS-SP2 Laser scanning confocal microscope with a 40X objective (Leica Microsystems, Bannockburn, Illinois, USA).

To generate numerical data we use a research tool with applicable algorithms, termed herein as the NASARadiationTrackImage©[1] model [11]. The model is used to simulate the whole nucleus in a Monte Carlo scheme. The model uses a random walk polymer representation of each chromosome constrained to within territories within a cell nucleus of a given volume and shape. Ion track structure is represented by a radial dose formalism [14]. The application of a stochastic track structure is being currently tested (not presented here). The algorithm for induction of DSBs is described in [11].

As an initial estimate, a one-to-one correspondence between foci and DSBs is assumed (even though questioned by some researchers), and, therefore, foci are assigned to DSBs as 3D balls of a finite diameter. The stochastic pattern of DSBs for each Monte Carlo realization is produced, which is highly non-random for heavy ions. Non-randomness (meaning non-Poisson statistics of objects in a given area) in DSB patterns is due to the fact that the majority of DSBs are aligned along the ion trajectory, with only a very small fraction occurring at distances larger than one micron, as determined by the physical track structure [15]. Fig. 1 demonstrates the spatial distribution of DSBs and associated foci. If the foci touch each other in the 2D equatorial plane of a cell nucleus, they are said to be linked. This projection onto a 2D plane is necessary to reflect the experimental technique, in which the maximum projection (MP) of cells and objects in cells is used. Such plane has a parallel orientation for the flat nucleus parallel to the beam, and a perpendicular orientation to the beam in a round nucleus and a flat nucleus perpendicular to the beam. An MP set of foci, in which each focus is linked to another member of the set, is called a focus chain.

[1] NASARadiationTrackImage refers to the model's ability to analyze and visualize, *via* DNA damage foci, DSB patterns induced by radiation tracks (copyright by USRA).

An alternative method for counting foci is done with a segmentation algorithm. The following flowchart briefly outlines how it works [16]:



Input morphological parameters for the maximum object size (volume or area), border complexity, and maximum lateral extent in *X*, *Y* or *Z* dimension.

Apply a global low threshold to image *I(x, y, z)* to remove obvious noise.

Do the first round of sorting to create interconnected sets of pixels (voxels) associated with each non-zero pixel (voxel) that satisfy the morphological parameters. This procedure creates so-called valid pixels (voxels).

Check all valid pixels (voxels) for nearest neighbors to combine them into final interconnected sets, which are the final objects (such as segmented foci).

3 Results

3.1 Foci Chains

In Fig. 1, Panel A, foci chains can be seen in an MP (maximum projection) experimental image obtained from high-LET irradiated fibroblasts. The simulated foci chains (Fig. 1, Panel B) are shown in a model generated image for high-LET Ti ions. The difference is that in Panel A, the image is two-dimensional and obtained by a

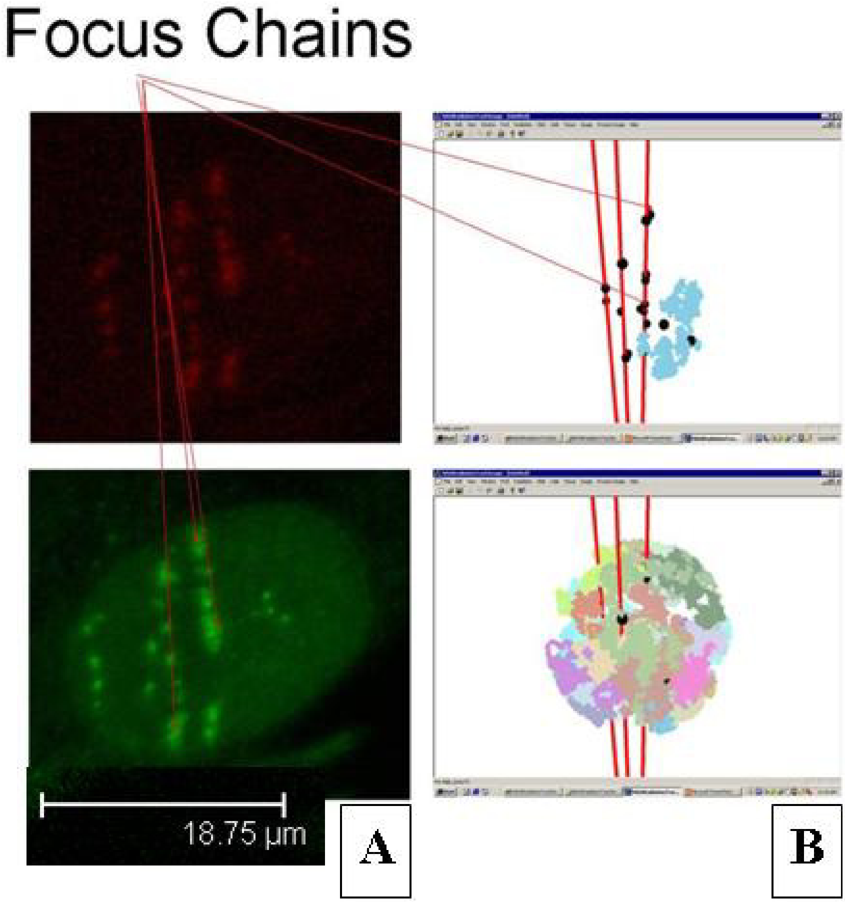


Fig. 1. Comparison of experimental images and model images of radiation-induced foci. Panel A. Formation of radiation-induced foci in noncycling normal human fibroblasts (HF19) exposed to 0.8 Gy Ti ions (E=1,000 MeV/n, LET=108 keV/µm). The simulated cell in Panel B received three hits of Ti ions. The scale in the insert below Panel A helps the viewer see the image scale and the size of the nucleus. Panel B. The GUI shows a simulated cell nucleus with the diameter 12 µm (which gives approximately the same scale for both Panels). Bright color fields in Panel B correspond to individual chromosomes, black balls visible inside the simulated nucleus are DSBs and red lines are heavy ion tracks.

maximum projection technique. In Panel B, the simulated foci are 3D objects; albeit they are simulated in a flattened nucleus and the foci chain are analyzed also in a projection to a 2D plane (parallel to the cell disk, or more precisely, the cell equator). In contrast to high LET, a simulation with low-LET X-rays produce foci, which are not linked in 99.4% of cases (data not shown).

To address whether the shape of the cell nucleus had any impact on the model calculations, we performed simulations to calculate the probability of the formation of a MP foci chain of a given length in a round nucleus and in a flat nucleus, both with the ion beam parallel and perpendicular to the plane of the cell monolayer. Flat nuclei were given by an ellipsoid with the same volume *V* as the round nucleus but with the principal axes (*a, b, c*) reduced by a factor of 4 in the *Z*-direction and extended by a factor of 2 in the lateral direction (for an ellipsoid, $V = abc = 904\ \mu m^3$). This orientation was perpendicular to the beam, a different manipulation of the parameters *a*, *b*, and *c* can make the nucleus disk to be parallel to the beam.

The MP foci chain probabilities (strictly, the P.D.F., or the probability density function for a chain of a given length to be realized) for the Fe, He, and Si ions are shown in Fig. 2. Herein, we demonstrate one example with Fe ions, *E*=1,000 MeV/n (LET=150 keV/μm) at *D*=1 Gy. More examples are given in [12], which include a fluence of exactly one ion per nucleus, He ions with *E*=0.75 MeV/n (corresponding to LET = 124 keV/μm), Si ions with *E*=90 MeV/n (corresponding to LET = 155 keV/μm). Using our model, we can vary ion type, LET and fluence to do sensitivity studies for the focus chain formation [12].

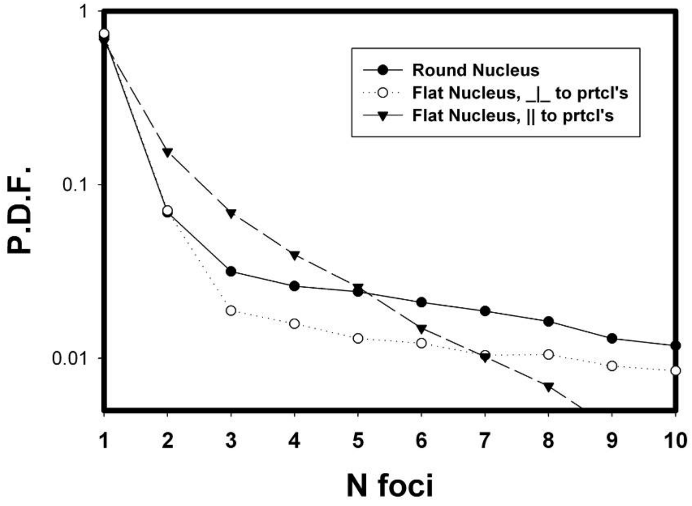


Fig. 2. The P.D.F. (probability density function) of MP foci chains *P(n)* vs # foci forming a chain. Three situations were simulated: nuclei were round as they might be in vivo; nuclei were flat, as they might be in a cell culture. For flat cells, the beam struck perpendicular or parallel to the plane of the cell monolayer in a cell culture. Three types of ions were considered: Fe ions, *E*=1,000 MeV/n (LET=150 keV/μm) at *D*=1 Gy.

The focus chain probabilities show some dependence on the shape of the nucleus. A flat nucleus oriented parallel to the beam appears to have a slightly higher probability of foci chains for a given *n* for shorter chains. Such orientation is what experimentalists observe usually. Even though the P.D.F. is higher for longer chain for the perpendicular orientation, one has to keep in mind that these are normalized distribution functions that have to cross over to preserve the area under the curve. Overall, the longer chains are very unlikely; therefore, the parallel orientation will have more visible chains from the experimentalist' point of view. The data show that the probability to have *n* foci in a chain even though drops off fast, remains non-zero for larger chains. The probability *P* for a focus to be contained within a chain of length 1, which means an isolated focus (not merged with other foci), was *P*(1)=0.684 for a round nucleus, *P*(1)=0.731 for a flat nucleus perpendicular to the beam, and *P*(1)=0.676 for a flat nucleus parallel to the beam (Fe ions, one ion per nucleus, first datum point in Fig. 2). In Fig. 2, *P*(1) shows not only how frequently a researcher would have to count merged foci (this frequency is given by 1-*P*(1)), but also gives a technique for adjusting the focus count, or focus yield. From the definition of the P.D.F., it follows that the total number of foci is the number of stand-alone foci divided by *P*(1). Higher *P*(1) observed for C (Fig. 3) and Si (Fig. 2, 3) ions indicate that the foci chains are less likely, as follows from the P.D.F. function definition. Fe ions, overall, are the worst in terms of inducing foci chains because of their higher focus yield per hit in addition to high *P*(1) and, therefore, the theoretical study is more useful for the Fe ion data. For example, in Fig. 3, for Fe ions, *P*(1)=0.671; for Si ions, *P*(1)=0.760; for C ions, *P*(1)=0.860, which means fewer stand-alone foci for Fe ions.

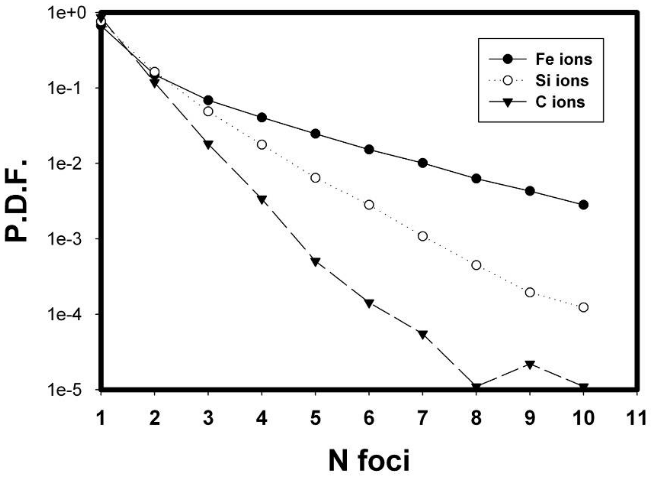


Fig. 3. Comparisons of the P.D.F. of MP chained foci vs. the length of the chain for a flat nucleus parallel to the incident radiation for three ions. Fe ions have a higher density of energy in the penumbra than Si and C ions, which is reflected in the higher likelihood of longer chains of foci (*D*=1 Gy, *E*=1,000 MeV/n for all ions). At this energy, $\mathrm{LET}_{\mathrm{Fe}}$= 150, $\mathrm{LET}_{\mathrm{Si}}$ = 43, $\mathrm{LET}_{\mathrm{C}}$ = 8 (keV/μm).

Other calculations are done to compare the impact of different ions on foci patterns, rather than the impact of the nucleus geometry. In Fig. 3, all nuclei are flat with the nucleus disk being parallel to the particle beam and all ions have energy E=1,000 MeV/n (but different LET's). Other LET's and E's are presented in [12]. The other data show that, at low energy E (very high LET), the lighter ions (C ions) are more likely to produce some chains, as their $P(1)$ was a little less than for the other ions.

3.2 Example of Image Analysis

We can use the image in Fig. 1 as a demonstration of the technique that uses $P(1)$, the probability to have a chain with one focus, or, simply, an isolated focus. In Fig. 1, Panel A, if one uses manual counting of the number of isolated islands of foci (foci chains), then the count is about 22. The model prediction, or the automated counting, predicts that the number of foci should be 22.0/0.697=31.6. To reiterate, this is the number of apparent foci divided by the model-predicted $P(1)$. Here and elsewhere the data are based on the MP algorithm that allows us comparing the data with the image in Fig. 1. Therefore, we predict that this particular nucleus has 31.6 DSBs.

We demonstrate the use of an additional tool that NASARadiationTrackImage GUI (graphics user interface) has. The presented segmentation procedure is not a model, but rather a technique that counts objects in images, which are foci in this case. This tool was developed as an attempt to automate image analysis and does not take into account track and chromatin geometry and statistical effects of DNA damage at different LET. The segmentation algorithm [16] identified 32 objects, which takes into account more subtle, not visible to human eye, variations in the intensity between neighboring foci (Fig. 4). In this example, 32 DSBs from segmentation is approximately equal to 31.6 DSBs from the above analysis, which shows that a different technique concurs with the model.

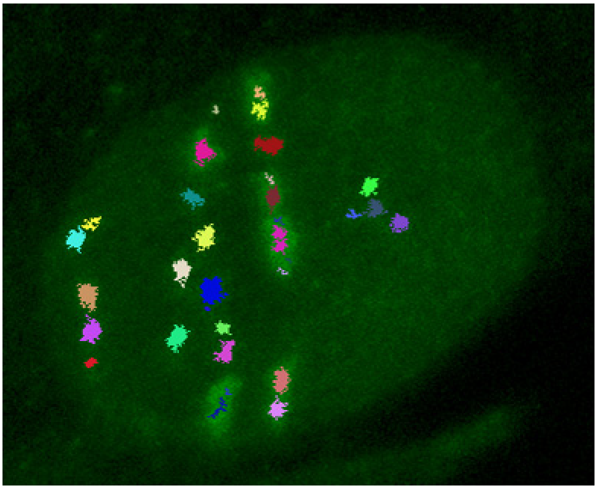


Fig. 4. The segmentation algorithm (part of NASARadiationTrackImage GUI) identified at least 32 DNA damage foci in the nucleus shown in Figure 1, Panel A.

4 Discussion and Conclusions

This paper addresses the difficulty of counting merged foci, which is used to evaluate the number of DSBs in irradiated cells. The proposed analysis is based on the assumption that the number of foci corresponds to the number of DSBs [17,18], even though this is not entirely correct [9,11,19]. The problem of the one-to-one correspondence of foci to DSBs is not addressed here, in this work we only considered the problem of overlapping or chained foci. The present analysis does not take into account the kinetics of foci evolution. Our analysis is good for a given time in the experiment. As experimental time progresses, the distribution of sizes of foci can change [13]. The analysis has not addressed this change in size, even though such analysis can be easily added.

Other important work in the area of foci counting that takes into account chromatin organization was done in [19, 20]. These efforts addressed the impact of euchromatin/heterochromatin on foci distributions in images [19], and the impact of chromatin remodeling and ensuing DSB relocation to DSB repair centers, as evidenced from foci images recorded in 3D, as well as time [20].

In radiation assays with high-LET particles, the DNA damage foci form streaks, and their density in the vicinity of the track center is sufficiently high to create optical complications for the count of foci. The other two complications include the multiple DSBs in a focus and the precise conversion of an isolated focus into a DSB (or lack of one-to-one correspondence between foci and DSBs).

A simple and straightforward way proposed to deal with the merged foci is a modification of the program, NASARadiationTrackImage [11], where the algorithm that simulates MP chained foci gives a better quantification of the merged foci. These simulations produce chained foci both graphically and numerically, and the numerical data include a statistics of focus chains of different lengths.

The presented approach offered a quick way to get a better focus and associated DSB yields. One can use the proposed P.D.F. for the focus chains and $P(1)$, the probability of a focus to be not chained. Using the definition of a P.D.F., the DSB yield is simply

$$DSB_{yield} = \frac{\overline{N}}{P(1)}, \tag{1}$$

where $\overline{N}$ is the number of not chained foci (and other values are defined above).

The difficulty to count the number of actual foci present in real images was resolved by the model analysis. Estimates on the probability of defined foci chains were calculated with a stochastic Monte Carlo algorithm that is based on the track structure and predicts the distribution of DNA damage. The analysis of foci chains will be useful for the experimentalists to give proper estimates of the dose-dependent DSB yield for high-LET particles.

Acknowledgments. The simulations were done on the NASA JSC Beowulf computer cluster machine. Funding was through the NASA Risk Assessment Project.

References

1. Cucinotta, F.A., Durante, M.: Cancer risk from exposure to galactic cosmic rays: Implications for space exploration by human beings. Lancet Oncology 7, 431–435 (2006)
2. Schulz-Ertner, D., Tsujii, H.: Particle radiation therapy using proton and heavier ion beams. Journal of Clinical Oncology 25, 953–964 (2007)
3. Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S., Bonner, W.M.: DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. Journal of Biological Chemistry 273, 5858–5868 (1998)
4. MacPhail, S.H., Banath, J.P., Yu, Y., Chu, E., Olive, P.L.: Cell cycle-dependent expression of phosphorylated histone H2AX: Reduced expression in unirradiated but not X-irradiated G1-phase cells. Radiation Research 159, 759–767 (2003)
5. Leatherbarrow, E.L., Harper, J.V., Cucinotta, F.A., O'Neill, P.: Induction and quantification of γ-H2AX foci formation following low- and high-LET radiation. International Journal of Radiation Biology 82, 111–118 (2006)
6. Desai, N., Davis, E., O'Neill, P., Durante, M., Cucinotta, F.A., Wu, H.: Immunofluorescence detection of clustered gamma-H2AX foci induced by HZE-particle radiation. Radiation Research 164(4), pt. 2, 518–522 (2005)
7. Han, J., Hendzel, M.J., Allalunis-Turner, J.: Quantitative analysis reveals asynchronous and more than DSB-associated histone H2AX phosphorylation after exposure to ionizing radiation. Radiation Research 165, 283–292 (2006)
8. Sedelnikova, O.A., Rogakou, E.P., Panyutin, I.G., Bonner, W.M.: Quantitative detection of (125)IdU-induced DNA double-strand breaks with gamma-H2AX antibody. Radiation Research 158(4), 486–492 (2002)
9. Cucinotta, F.A., Pluth, J.M., Anderson, J., Harper, J.V., O'Neill, P.: Biochemical Kinetics Model of DSB Repair and gamma-H2AX Foci by Non-Homologous End Joining. Radiation Research 169, 214–222 (2008)
10. Furuta, T., Takemura, H., Liao, Z.Y., Aune, G.J., Redon, C., Sedelnikova, O.A., Pilch, D.R., Rogakou, E.P., Celeste, A., Chen, H.T., Nussenzweig, A., Aladjem, M.I., Bonner, W.M., Pommier, Y.: Phosphorylation of Histone H2AX and Activation of Mre11, Rad50, and Nbs1 in Response to Replication-dependent DNA Double-strand Breaks Induced by Mammalian DNA Topoisomerase I Cleavage Complexes. The Journal of Biological Chemistry 278, 20303–20312 (2003)
11. Ponomarev, A.L., Costes, S.V., Cucinotta, F.A.: Stochastic properties of radiation induced DSBs: DSB distributions in large scale chromatin loops, the HPRT gene and within the visible volumes of DNA repair foci. International Journal of Radiation Biology 84(11), 916–929 (2008)
12. Ponomarev, A.L., Huff, J., Cucinotta, F.A.: The Analysis of the Densely Populated Patterns of Radiation-Induced Foci by a Stochastic, Monte Carlo Model of DNA Double-Strand Breaks Induction by Heavy Ions. International Journal of Radiation Biology 86(6), 507–515 (2010)
13. Chappell, L.J., Whalen, M.K., Gurai, S., Ponomarev, A.L., Cucinotta, F.A., Pluth, J.M.: Analysis of Flow Cytometry DNA Damage Response Protein Activation Kinetics after Exposure to X rays and High-Energy Iron Nuclei. Radiation Research 174(6a), 691–702 (2010)
14. Cucinotta, F.A., Nikjoo, H., Goodhead, D.T.: Comment on the effects of delta-rays on the number of particle-track transversals per cell in laboratory and space exposures. Radiation Research 150, 115–119 (1998)

15. Cucinotta, F.A., Nikjoo, H., Goodhead, D.T.: Model of the Radial Distribution of Energy Imparted in Nanometer Volumes from HZE Particle. Radiation Research 153, 459–468 (2000)
16. Ponomarev, A.L., Davis, R.L.: An Adjustable-Threshold Algorithm for the Identification of Objects in Three-Dimensional Images. Bioinformatics 19, 1431–1435 (2003)
17. Rothkamm, K., Löbrich, M.: Evidence for a lack of DNA double-strand break repair in human cells exposed to very low x-ray doses. Proceedings of the National Academy of Sciences USA 100, 5057–5062 (2003)
18. Stiff, T., O'Driscoll, M., Rief, N., Iwabuchi, K., Löbrich, M., Jeggo, P.: ATM and DNA-PK Function Redundantly to Phosphorylate H2AX after Exposure to Ionizing Radiation. Cancer Research 64, 2390–2396 (2004)
19. Costes, S., Ponomarev, A.L., Chen, J., Cucinotta, F.A., Barcellos-Hoff, M.H.: Chromosome model reveals dynamic redistribution of DNA damage in nuclear sub-domains. PLoS Computational Biology 3(8), e155 (2007)
20. Jakob, B., Splinter, J., Conrad, S., Voss, K.-O., Zink, D., Durante, M., Löbrich, M., Taucher-Scholz, G.: DNA double-strand breaks in heterochromatin elicit fast repair protein recruitment, histone H2AX phosphorylation and relocation to euchromatin. Nucleic Acids Research 39(15), 6489–6499

Multiple CPU Computing: The Example of the Code RITRACKS

Ianik Plante[1,2,*] and Francis A. Cucinotta[2]

[1] Division of Space Life Sciences, Universities Space Research Association,
3600 Bay Area Boulevard, Houston, TX 77058
Ianik.Plante-1@nasa.gov

[2] NASA Johnson Space Center, Bldg 37, Mail Code SK, 2101 NASA Parkway, Houston,
TX 77058

Abstract. The Monte-Carlo simulation code RITRACKS is used to simulate the radiation track structure of heavy ions and electrons. The original version of the code RITRACKS was developed to perform calculations using one central process unit (CPU). New computers now comprises several cores for computation, allowing the processing of several instructions simultaneously and of parallel computing. While computers based on multi-core CPU offers potentially tremendous improvement in terms of performance, the computer programs which were originally developed to work on a single CPU needs to be adapted to use multiple cores. In this paper, we discuss how the code RITRACKS was modified and adapted to use multiple CPU on a Windows-based workstation by using the TThread object provided in the CodeGear(TM) RAD Studio development environment. The advantages and limitations of this approach will also be discussed.

Keywords: Multiple CPU computing, Monte-Carlo simulations, radiation track structure, C++, TThread object.

1 Introduction

The simulation of the radiation track structure is of great interest in several fields, such as in radiotherapy treatment planning, micro and nanodosimetry, space radiation, radiation chemistry and DNA damage studies [1]. Amorphous track structure models have been used for several decades to understand the initial interactions of ions with matter [2]. The first Monte-Carlo code track structure simulations was developed later, in the 1980's [3]. Since then, many other simulation codes have been developed, with different purposes (reviewed in [1]). Because of the stochastic nature of ionizing radiation interactions, Monte-Carlo simulations provide a better representation of the track structure, which is not described by amorphous track models [4]. However, Monte-Carlo codes are more difficult to use than amorphous track models because they require much more computation power and time as well as a significant amount

[*] Corresponding author.

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 12–25, 2013.

of storage space due to the large number of histories which is often necessary to minimize statistical fluctuations on the results.

The software RITRACKS (Relativistic Ion Tracks) is a Monte-Carlo program used to simulate radiation tracks for heavy ions and electrons [4-6]. In principle any ion track can be simulated if the energy is within the range of which the cross sections for interactions of primary particles and secondary electrons with target molecules are known. The dose deposited by the radiation can be calculated in microvolumes [7] and in nanovolumes (voxels) [8]. An example of a track structure calculated by RITRACKS is shown in Figure 1.

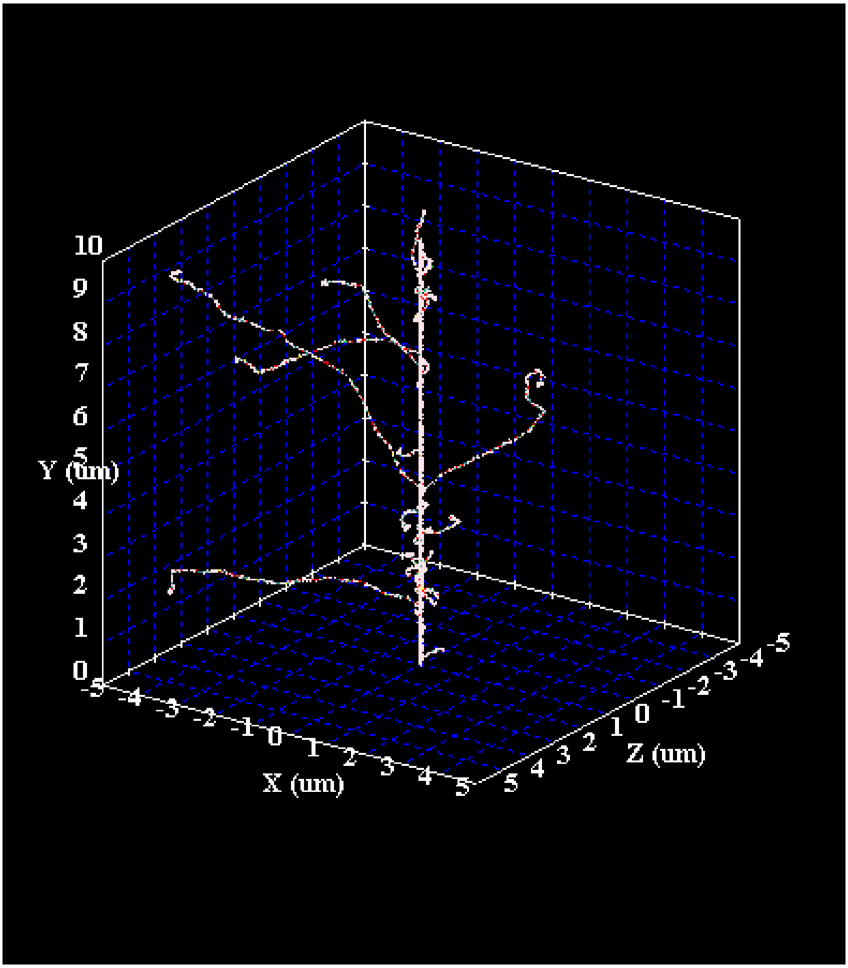


Fig. 1. Visualization of the track structure of a $^{12}C^{6+}$ ion, 25 MeV/u, as simulated by RITRACKS (linear energy transfer ~78 keV/µm)

The code has been validated with experimental data. For instance, RITRACKS was used recently to simulate radial dose experiments of Schmollack and coworkers [9] in 100 nm diameter volumes. Most experimental data such as the frequency of hits and the mean energy deposited in the detector per event and per ion were reproduced satisfactorily by these simulations.

The drawback of the use of RITRACKS is its long calculation time. Regarding this aspect, the use of computers with multiple CPU can significantly reduce the calculation time for a given task and/or allow more tasks to be completed in a given time. In this paper, a method developed to use several CPUs on Microsoft Windows-based systems is explained in details. This method uses the TThread object provided in the CodeGear(TM) RAD Studio C++ Builder environment. By using this method, we have been able to decrease the simulation time necessary for RITRACKS to perform a given task. The program presented here can be easily adapted to parallelize similar codes. Finally, simulation performance results are given and issues with this approach is discussed.

2 Materials and Methods

The code RITRACKS was developed using the CodeGear(TM) RAD Studio environment in the programming language C++ and is designed to work on Windows systems. The major components of RITRACKS are 1) The differential and integrated interaction cross sections of ions and electrons with water; 2) Cross sections sampling algorithms and routines; 3) Particles transport routines; 4) Data collection and management (notably calculation of dose); 5) Input/output routines; 6) A graphic user interface (GUI); 7) A 3D interface for the visualization of the track structure; 8) Cross section visualization windows; 9) A help file; 10) Redistributable libraries. The calculation part "TrackCalc.exe" is an independent program which is compiled separately from the GUI. This part can also be compiled and executed on Linux machines. The calculation part of the code RITRACKS calculates the energy deposition events, ionization and excitation of water molecules by the heavy ion and the energy, the position and direction of the secondary electrons, as well as the tracks of the secondary electrons. The ion and electron cross sections, transport methods and simulation results were given in our previous publications [4-6]. The algorithm of the calculation part is summarized in Figure 2.

The computer that was used for this work is a Dell Workstation comprising a Intel(R) Xeon(R) CPU E5430 @ 2.66 GHz. This computer is a 4-core using the hyper-threading technology, that is, for each processor core that is physically present, the operating system addresses two virtual processors, and shares the workload between them when possible. The technology is transparent to operating systems and programs. Therefore, the 4-core machine behaves as a 8-cores machine.

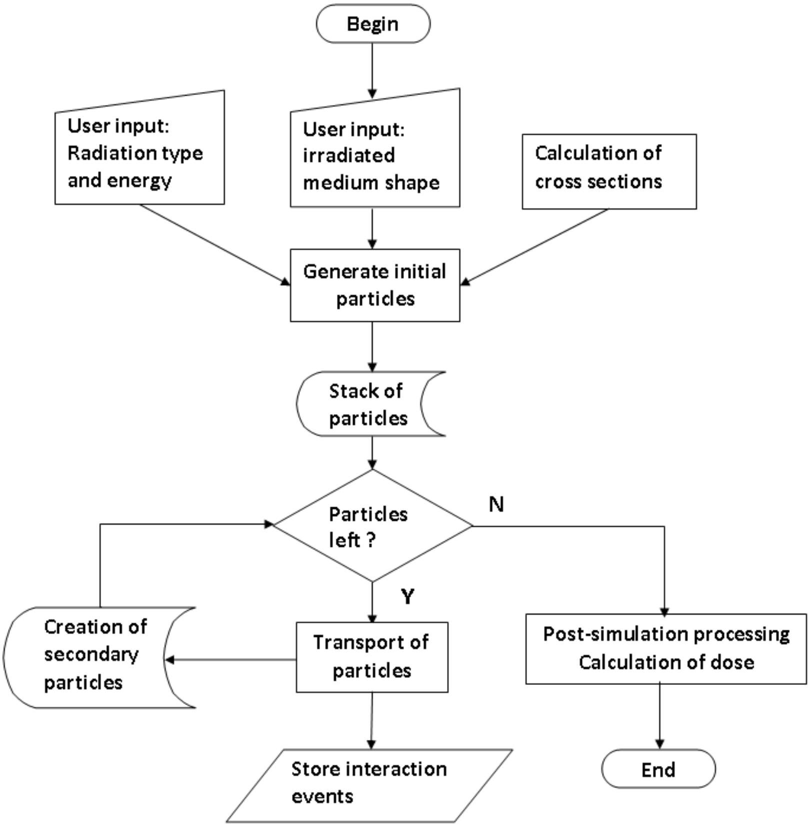

Begin
User input: Radiation type and energy
User input: irradiated medium shape
Calculation of cross sections
Generate initial particles
Stack of particles
Particles left ?
N
Y
Creation of secondary particles
Transport of particles
Post-simulation processing Calculation of dose
Store interaction events
End

Fig. 2. The algorithm of code RITRACKS

The steps to use the full features of multiple CPU machines are: 1) Copy of the program and necessary input and data files in separate directories; 2) Execution of tasks on available CPUs; 3) Collection of results. This is illustrated in Figure 3.

2.1 Copy of the Program and Necessary Files

The initial condition file is also created at this time. In each copy of this file, an initial SEED value to generate a particular sequence of random numbers is given [10]. Therefore, each copy of the program is the same, but is executed by using a different sequence of random numbers. Consequently, different tracks are generated by each copy of the program, because these random numbers are used as described in [4] to generate tracks. Therefore, an average can be made over a large number of tracks in a short time.

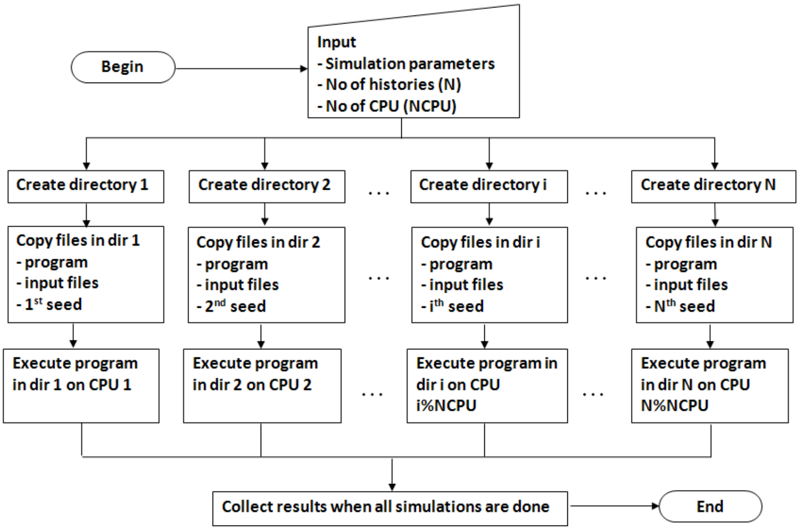


Fig. 3. The use of multiple CPU

2.2 Execution of Tasks on Available CPUs

This section is the most important for the use of multiple CPU. In Borland C++, this is done by using the object TThread. These objects are used to control the execution of the program. Their structure is as follows:

File Unit2.cpp

```
#include <vcl.h>
#include "Unit2.h"

__fastcall
ThreadRunExtProgram::ThreadRunExtProgram(bool
CreateSuspended, int tNo, int tCPUNo, String wDir,
String eName, bool sConsole):
TThread(CreateSuspended)
{
    threadNo                  = tNo;
    threadCPUNo               = tCPUNo;
    workingDirectory          = wDir;
    exeName                   = eName;
    showConsole               = sConsole;
    extThreadNo               = threadNo;
    dirToThread               = wDir
    +"\\Copy"+String(threadNo).c_str();
}
```

```
void __fastcall ThreadRunExtProgram::Execute()
{
    lpCurrentDirectory     = workingDirectory+"/"
                           +"Copy"+String(threadNo)
                           +"/";
    szCmdLine1           = workingDirectory
                           +"/"+"Copy"
                           +String(threadNo)+"/"
                           +exeName;

    // Execute external application
    privateErrorCode       =
ExecuteExt(threadCPUNo,szCmdLine1.t_str(),
  lpCurrentDirectory.t_str(), showConsole);

    Synchronize(AffMessage);
}

void __fastcall ThreadRunExtProgram::AffMessage()
{
    errorCodeThread        = privateErrorCode;
}
```

The corresponding header file is

File Unit2.h

```
#ifndef Unit2H
#define Unit2H
//-----------------------------------------------------------
--------------------
#include <Classes.hpp>
//-----------------------------------------------------------
--------------------
class ThreadRunExtProgram : public TThread
{
  private:

    String   szCmdLine1;
    String   lpCurrentDirectory;
    int      privateErrorCode;
    int      threadNo;
    int      threadCPUNo;
    String    workingDirectory;
    String    exeName;
    bool      showConsole;

protected:
    void __fastcall Execute();
    void __fastcall AffMessage();

public:
    int __thread errorCodeThread;
    int __thread extThreadNo;

    __fastcall ThreadRunExtProgram(bool
CreateSuspended, int tNo, int tCPUNo,
    String wDir, String eName, bool sConsole);
};
//----------------------------------------------------------
--------------------
#endif
```

The main parts are the creation (ThreadRunExtProgram(...)), the execution (Execute()) and synchronization with variables which do not belong to the thread (AffMessage()). Most variables used in Unit2.cpp and Unit2.h have self-explanatory names. The objects ThreadRunExtProgram are declared as follows in the main program:

```
ThreadRunExtProgram **threads = new
ThreadRunExtProgram*[nMaxThreads];
```

where nMaxThreads is the max number of threads, set arbitrarily to 10000. They are initialized as follows:

```
for (int i=0; i < noThreads; i++) {

  threads[i]        = new ThreadRunExtProgram(true,
                    i, i%nCPUUsed,
                    workDirectories(i).c_str(),
                    "TrackCalc.exe", showConsole);
  threads[i]->FreeOnTerminate      = true;
  threads[i]->OnTerminate      = EndOfThread;
}
```

The variable CreateSuspended is set to true, to prevent the threads from being executed at this moment. A CPU number i%nCPUUsed is assigned to the task, since it is possible to execute the program "TrackCalc.exe" on a specific CPU by using the routine given in Appendix I. The function EndOfThread() is called when the thread is done. The instruction WaitForSingleObject() in the routine ExecuteExt() is particularly important, because the thread is halted at this point and waits for the program "TrackCalc.exe" to complete or for a run-time error to occur.

The threads are executed as follows:

```
for (int i=0; i < noThreads && i<nCPUUsed; i++) {
  threads[i]->Resume();
}
```

The variable nCPUUsed indicates the number of CPUs which should be used for the simulation. It is possible to execute more tasks than the number of CPUs simultaneously, but the program limits the nCPUUsed to the total number of CPUs to avoid more than one task running simultaneously on the same CPU. The routine EndOfThreads() is called whenever a task is completed.

```
void __fastcall TForm1::EndOfThread(TObject *Sender)
{
  ThreadRunExtProgram *threadTemp    = dynamic_cast
<ThreadRunExtProgram   *>(Sender);
  int nT                             = threadTemp
  ->extThreadNo;

  if (threadTemp->errorCodeThread==0) {
    statusLabels[nT]->Font->Color    = clBlack;
    statusLabels[nT]->Caption        = "Simulation
                                        complete";
    completeExecution[nT]            = true;
  }
  else
  {
    statusLabels[nT]->Font->Color    = clRed;
    statusLabels[nT]->Caption        = "This
                                       simulation
                                       crashed!!";
    crashedTasks++;
  }

  completedTasks++;

  int noNext                         = nT+nCPUUsed;

  if (noNext<noThreads) {
    threads[noNext]->Resume();
    statusLabels[noNext]->Caption    = "Starting
                                       simulation";
    Form6->Update();
  }
  if (completedTasks==noThreads)
  {
    Form1->RichEdit1->Lines->Append("All threads
    complete. Collecting results.");
    PostSimulation();
  }
}
```

The instance of ThreadRunExtProgram which called the routine EndOfThreads() is obtained by using the function dynamic_cast. The routine EndOfThreads() then verifies if an error code is returned, which indicate that there was a run-time error in the execution of the program "TrackCalc.exe". Then the next task is executed.

2.3 Collection of Results

When all the tasks are completed, the routine PostSimulation() is called. The results from the different simulations are compiled and averaged in this routine. An important aspect to consider regarding the collection of results is the synchronization of the threads. To optimize the CPU usage, all processors should be assigned a similar task. By using the approach described in this paper, since the only difference between threads is the sequence of random numbers, all threads are expected to complete more or less at the same time.

3 Results

The simulation time to perform the simulation of 50 $^{12}C^{6+}$ of 25 keV/μm tracks for a given number of CPUs is given in Figure 4.

The number of CPUs in use can be verified by looking at the Performance tab of the Windows Task Manager. A sharp rise in the CPU usage is usually seen on the active CPUs. The simulation time decreases by from 529 s to 282 s by using two CPUs instead of one. Then the simulation time decreases to a lesser extent to reach a plateau at ~200 s when four or more CPUs are used. Similar studies of the simulation time for a given task as a function of the number of CPUs used were performed (not shown); similar results regarding the calculation time were found. The fact that there is no improvement of the calculation time by using over four CPUs may indicate that our method do not use the full advantages of the hyperthreading technology.

We have also calculated the number of tracks a per second function of the number of CPU. as shown in Figure 5. To perform this study, each processor is assigned the task of simulating 20 tracks.

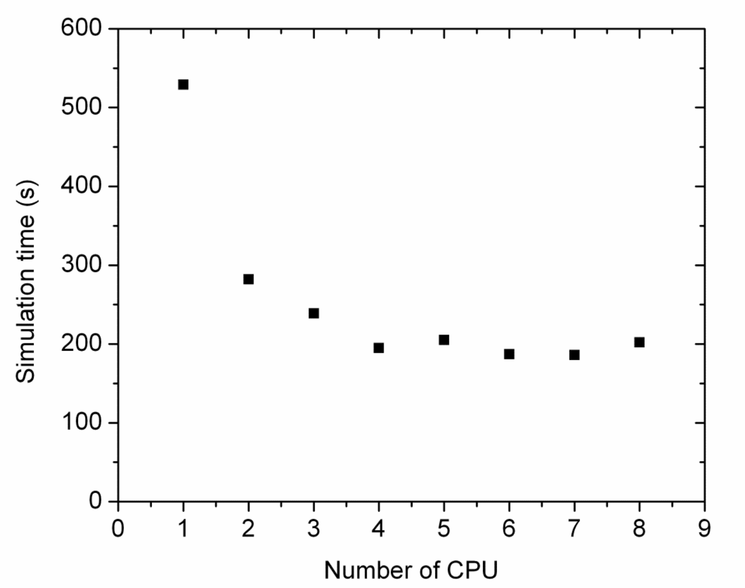


Fig. 4. Simulation time as a function of the number of CPU for the simulation of 50 $^{12}C^{6+}$ ion tracks, 25 MeV/amu

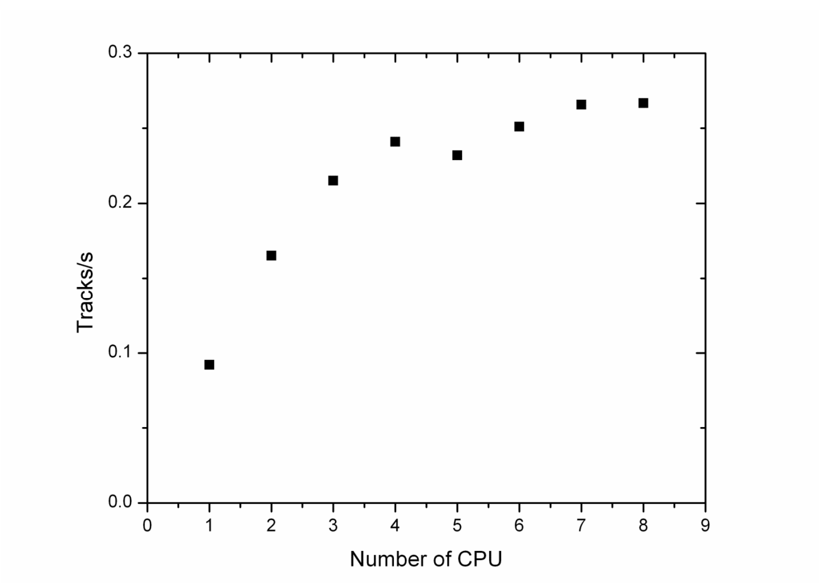


Fig. 5. Number of calculated tracks per second as a function of the number of CPU for $^{12}C^{6+}$ ion tracks, 25 MeV/amu. Each CPU is assigned the task of simulating 20 tracks.

The number of calculated tracks per second increase from ~0.09 tracks/s for 1 CPU to ~0.24 tracks/s for 4 CPU. The number of calculated tracks per second do not increase much by further increasing the number of CPU.

4 Discussion

Processors comprising several CPUs are now routinely used in today's computers, and there is a trend for manufacturers to increase the number of CPUs present in each machine. These processors have the potential to greatly increase performance of programs, but it is necessary to parallelize the codes to use their full potential. In this paper, we have presented the improvement in simulation time by using multiple CPU for the code RITRACKS. Although the gain of performance is relatively modest, there are many good reasons to use this approach. First, a gain of a factor 2-3 may means that several hours or days of calculations can be saved. Second, the parallelization code presented here is expected to yield further gain of performance on newer systems comprising more cores. Third, when a long simulation is done by a conventional program using one CPU, it may crash. In most cases, the data is usually lost. The simulation using multiple CPU can be useful, because the final results are obtained by averaging the data from successful tasks only. Therefore, it may also help the software development by tracking errors, because the same program can be executed simultaneously using different initial conditions, notably a different sequence of random numbers.

An issue possibly limiting performance with multicore computers may be the input/output (I/O), notably writing to disk. If several processes are writing to the same disk simultaneously, it may significantly slow down the whole task. Even if the disk is partitioned, the processes are writing to the same physical disk. Therefore it may be useful to have several physical hard drives on a computer comprising multiple cores. It would be possible to assign a disk to a task. Another issue that was encountered regarding I/O was copying of files, which is done automatically by the program before and after the threads. We have found that the conventional built-in Microsoft Windows "copy" and "xcopy" commands are not thread-safe, and that some files were not copied correctly. The situation was improved by using the command "Robocopy" included in the Microsoft Windows Resource Kit, but not fully corrected. Therefore we have added the option to verify if a copy is done properly after an attempt.

Parallelizing a computer code is not a simple task, and there are probably several ways for doing it. In our case, the threads are independent from each other; therefore the approach presented in this paper is relatively simple and may also be used for similar programs, notably Monte-Carlo simulation codes which usually require a large number of histories. The method presented here can be easily modified to execute other programs using multiple CPU, allowing faster computational time and/or use of larger number of histories. Another technology, the general purposes graphic processing units (GPGPU) have appeared recently. These are improved graphic cards comprising multiple cores with a simpler architecture than normal CPU, and which can be used to speed-up parallel calculations. Unfortunately, Monte-Carlo track structure codes are too complex to use with available GPGPU.

Acknowledgements. This work was supported by the NASA Space Radiation Risk Assessment Project. Some parts of the code were adapted from examples posted on the web sites www.chgi.developpez.com and www.icrontic.com.

References

1. Nikjoo, H., Uehara, S., Emfietzoglou, D., Cucinotta, F.A.: Track-Structure Codes in Radiation Research. Radiat. Meas. 41, 1052–1074 (2006)
2. Cucinotta, F.A., Katz, R., Wilson, J.W.: Radial Distribution of Electron Spectra From High-Energy Ions. Radiat. Env. Biophys. 37, 259–265 (1998)
3. Turner, J.E., Hamm, R.N., Wright, H.A., Ritchie, R.H., Magee, J.L., Chatterjee, A., Bolch, W.E.: Studies to Link the Basic Radiation Physics and Chemistry of Liquid Water. Radiat. Phys. Chem. 32, 503–510 (1988)
4. Plante, I., Cucinotta, F.A.: Monte-Carlo Simulation of Ionizing Radiation Tracks. In: Mode, C.B. (ed.) Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science. InTech, Rijeka (2011)
5. Plante, I., Cucinotta, F.A.: Ionization and Excitation Cross Sections for the Interaction of HZE particles and Application to Monte-Carlo Simulation of Radiation Tracks. N. J. Phys. 10, 125020 (2008)
6. Plante, I., Cucinotta, F.A.: Cross Sections for the Interactions of 1 eV-100 MeV Electrons in Liquid Water and Application to Monte-Carlo Simulation of HZE Radiation Tracks. N. J. Phys. 11, 063047 (2009)
7. Plante, I., Cucinotta, F.A.: Energy Deposition and Relative Frequency of Hits of Cylindrical Nanovolume in Medium Irradiated by Ions: Monte-Carlo Simulations of Track Structure. Radiat. Env. Biophys. 49, 5–13 (2010)
8. Plante, I., Ponomarev, A.L., Cucinotta, F.A.: 3D Visualization of the Stochastic Patterns of the Radial Dose in Nano-Volumes by a Monte-Carlo Simulation of HZE Ion Track Structure. Radiat. Prot. Dosim. 143, 156–161 (2011)
9. Schmollack, J.U., Klaumuenzer, S.L., Kiefer, J.: Stochastic Radial Dose Distributions and Track Structure Theory. Radiat. Res. 153, 469–478 (2000)
10. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in Fortran, 2nd edn. Cambridge University Press, Cambridge (1992)

Appendix I: Routine to Execute a Program on a Specific CPU

The following code is used to execute a program on a specific CPU, and wait for it to complete.

```
#include <windows.h>
#include <strsafe.h>

int ExecuteExt(int CPUNo, String processName,
LPCTSTR lpCurrentDirectory, bool showConsole )
{
  DWORD       dwExitStatus;
  STARTUPINFO    si               = {sizeof(si)};
  PROCESS_INFORMATION  pi         = {0};
  char        *pszArgs;
  ULONG       ulValue;
  DWORD_PTR   dwProcess, dwSystem;
  DWORD       dwCreationFlags;
  DWORD       dwExitCode          = 0;
  DWORD       RetCode;
  String      sCPUNo              = CPUNo;
  String      space               = " ";

  String      sCmdLine            =
              sCPUNo+space+processName;
  PSTR        szCmdLine           =
              sCmdLine.t_str();
  TCHAR*      appName             =
              processName.t_str();

  if (showConsole) {
    dwCreationFlags =
    CREATE_SUSPENDED|NORMAL_PRIORITY_CLASS;
  }
  else
  {
    dwCreationFlags =
    CREATE_SUSPENDED|CREATE_NO_WINDOW;
  }

  pszArgs                         = szCmdLine;

  // Make sure we got a valid cpu index
  ulValue = 1 << strtoul((const char *)
  &szCmdLine[0], &pszArgs, 10);
  GetProcessAffinityMask(GetCurrentProcess(),
  &dwProcess, &dwSystem);
  ulValue &= ulValue & dwSystem;
  if (!ulValue || pszArgs == &szCmdLine[0])
  {
    MessageBox(NULL, "cpcpu <CPU n> <command
    line>\r\n\r\nn = 0 based CPU index", "USAGE",
    MB_ICONINFORMATION);
    return ~ERROR_SUCCESS;
  }

  while(*pszArgs && isspace(*pszArgs)) pszArgs++;


  if (*pszArgs)
  {
    CreateProcess(NULL, pszArgs, NULL,NULL,FALSE,
    dwCreationFlags, NULL,lpCurrentDirectory,&si,
    &pi);

    // Handle from CreateProcess is
       PROCESS_ALL_ACCESS
    if (pi.hProcess)
    {
      // Make it run where we want
      SetProcessAffinityMask(pi.hProcess, ulValue);
      ResumeThread(pi.hThread);

      // Wait for the process to finish
      WaitForSingleObject(pi.hProcess,INFINITE);

      GetExitCodeProcess( pi.hProcess, &RetCode );

      CloseHandle(pi.hThread);
      CloseHandle(pi.hProcess);
    }
    else
    {
      int er = GetLastError();
      return er;
    }
  }
  else
  {
    MessageBox(NULL, "cpcpu <CPU n> <command
    line>\r\n\r\nn = 0 based CPU index","USAGE",
    MB_ICONINFORMATION);
    return ~ERROR_SUCCESS;
  }
  return RetCode;
}
```

Modeling the Depressed Hematopoietic Cells for Immune System under Chronic Radiation

Shaowen Hu[1] and Francis A. Cucinotta[2]

[1] Universities Space Research Association, Division of Space Life Sciences,
Houston, Texas 77058, USA
shaowen.hu-1@nasa.gov
[2] NASA, Lyndon B. Johnson Space Center, Houston, Texas 77058, USA
francis.a.cucinotta@nasa.gov

Abstract. Although moderate dose (0.5 to 2 Gy) of ionizing radiation (IR) is well recognized to cause various disorders of the hematopoietic system (e.g., short-term effects like cytopenia, and long-term effects like leukemia), many quantitative aspects of the dynamics of the hematopoiesis response to long duration low dose rate IR still require additional investigation. Recently two cell kinetics models after acute radiation exposure are proposed to describe the perturbation of granulocytes and lymphocytes, respectively, in peripheral blood of various mammals. These two models are indeed built on a similar coarse-grained structure of hematopoietic system, thus they have the potential to form a unified model to characterize the mammalian hematopoietic system after various types of IR exposure. In this study we investigate the capability of the models to simulate the data of hematological measurements of the Techa River residents chronically exposed to IR in 1950-1956. Our modeling investigation indicates human hematopoietic precursor cells are more sensitive to chronic radiation than previously considered.

Keywords: biomathematical model, hematopoiesis system, radiation effect, chronic radiation.

1 Introduction

Space radiation is one of the many unavoidable factors that may cause serious health hazards to astronauts in space exploration. Beside the unpredictable sporadic solar particle events (SPE) which can impose relative large dose of radiation in a short timeframe, low level chronic radiation from Galactic Cosmic Rays (GCR) with high-energy and charge particles is another major concern due to the particles' high penetrating power to deep seated organs and their high radiation quality factors [6, 7, 8]. Among many tissues/organs of the human body that are sensitive to IR, the hematopoietic system is one of the most vulnerable. It has been established that the hematopoietic stem cell pool is very radiosensitive, even low level of IR exposure can cause perturbation in its proliferating function as well as its own maintenance [19]. This pool is responsible for the maintenance in the peripheral blood of different blood

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 26–36, 2013.

cell types (erythrocytes, granulocytes, platelets, lymphocytes, etc.), which perform many important functions such as oxygenation, nutrient and waste transportation, defense and immune response, coagulation, etc. Appraise the risk of human hematopoietic stem cell and hematopoietic system under various scenario of space radiation is one of the most important task for future space radiation protection [16].

Recently we have investigated two mathematical models of blood cell kinetics after radiation, based on a coarse-grained hematopoietic scheme proposed by Smirnova et al. [15, 18, 20]. One is for the granulopoiesis system, which describes the dynamical interactions of neutrophiles, eosinophiles, and basophiles in peripheral blood and their progenitor compartments in bone marrow. By utilizing species-dependent hematopoietic and radiobiological parameters for beagle dog, rhesus monkey, and human, this model can generate results consistent with experiments and empirical records from various sources, involving acute, protracted, and chronic radiation [12, 13]. Another model is for the lymphopoiesis system, which analyzes the lymphocyte changes in the blood of exposed human victims in radiation accidents [14]. Model simulations with reported absorbed doses as inputs can qualitatively and quantitatively describe a wide range of accidental data in vastly different scenarios. In addition, the absolute lymphocyte counts and the depletion rate constants calculated by this model show good correlation with two widely recognized empirical methods for early dose assessment [14]. These works demonstrate the potential to use Smirnova's models to build up a unified model to characterize mammalian hematopoietic response to IR.

Granulocytes and lymphocytes are the main components of mammalian immune system. The neutrophiles, the dominant type of granulocyte in all mammalians, are highly phagocytic and can kill a variety of microorganisms. The eosinophiles and basophiles are specialized to participate in allergic inflammatory responses [3]. Lymphocytes comprise of T cells, B cells and natural killer (NK) cells. NK cells are a part of the innate immune system and play a major role in defending the host from both tumors and virally infected cells. T cells are involved in cell-mediated immunity, whereas B cells are primarily responsible for humoral immunity (relating to antibodies). Though these peripheral cells have different radiosensitivity, all originate from a same hematopoietic stem cell pool, and all show certain degree of depression after acute, protracted, and chronic radiation [13, 14].

Previous study on beagle dogs indicates the chronic radiation effects of granulocytes can be modeled with a same scheme as the acute radiation but with a dose-rate dependent radiosensitivity parameter for bone marrow precursor cells [13]. Recently a series of results were published for the hematological measurements of the Techa River residents chronically exposed to IR in 1950-1956 [1, 2]. This prompts us to investigate the dose-rate dependent relationship for human hematopoietic system, so that the canine model can be reasonably extended to human model. In this study, based upon the suppressed levels of granulocytes and lymphocytes under different levels of chronic radiation, we find a unified relationship between the chronic dose-rate and radiosensitivity parameters can be identified, both for the granulopoiesis and lymphopoiesis models. This reinforces the modeling power of Smirnova's scheme on the radiation effects on mammalian hematopoiesis systems.

2 Mathematical Models

Based on the hematopoietic scheme of Smirnova [15], each hematopoietic line considers three coarse-grained compartments according to the degree of the maturity and differentiation of the cells:

- X_1, the bone-marrow precursor cells (from stem cells in the respective microenvironment to morphologically identifiable dividing cells);
- X_2, the nondividing maturing bone-marrow cells;
- X_3, the mature cells in peripheral blood.

The granulopoiesis model considers also the mature granulocytes in various tissues (X_4), as granulocytes only transiently stay in blood (usually a few hours), then migrate into tissues in an age-independent manner.

As nondividing maturing-only cells and mature cells for granulopoietic line are much more radio-resistant [11], only X_1 cells in granulopoiesis are assumed to be affected by radiation. Under whole body irradiation at a dose rate N, the dynamics of the concentration of four cell compartments is

$$\frac{dx_1}{dt} = Bx_1 - \gamma x_1 - \frac{N}{D_c} x_1, \tag{1}$$

$$\frac{dx_2}{dt} = \gamma x_1 - \delta x_2, \tag{2}$$

$$\frac{dx_3}{dt} = \delta x_2 - \kappa x_3, \tag{3}$$

$$\frac{dx_4}{dt} = \kappa x_3 - \psi x_4, \tag{4}$$

where x_i (i=1-4) are the concentrations of cells in compartment i, B is the reproduction rate of X_1 cells, γ, δ, and κ are the specific rates of transfer of cells from the various pools to the next pools, ψ is the decay rate of granulocytes in tissues, and D_c is radiosensitivity parameter which characterizes the capability of X_1 cells to maintain intact from any radiation damage. The dynamics of the damaged cells are described by

$$\frac{dx_{wd1}}{dt} = \left(\frac{N}{D_c} - \frac{N}{D_1} \right) x_1 + Bx_{wd1} - \gamma x_{wd1} - \nu_c x_{wd1}, \tag{5}$$

$$\frac{dx_{d1}}{dt} = \frac{N}{D_1} \frac{1}{1+\rho_1} x_1 - \nu_1 x_{d1}, \tag{6}$$

$$\frac{dx_{hd1}}{dt}=\frac{N}{D_1}\frac{\rho_1}{1+\rho_1}x_1-\nu_2 x_{hd1}, \tag{7}$$

where x_{wd1}, x_{d1}, and x_{hd1} are the concentrations of weakly damaged, damaged cells, and heavily damaged X_1 cells, respectively, D_1 is the conventional radiobiological dose D_0 for X_1 cells, i.e., a dose after which the cells in this compartment lose 63% of their initial number, ρ_1 the ratio of the numbers of X_{d1} and X_{hd1} cells, and ν_c, ν_1 and ν_2 the specific death rates of three types of damaged cells.

Based on the implicit regulation mechanism, the production rate of X_1 cells is determined by other parameters and cell concentrations:

$$B=\alpha/\{1+\beta[\theta_1(x_1+x_{wd1}+\Phi x_{d1}+\Gamma x_{hd1})+\theta_2 x_2+\theta_3 x_3+\theta_4 x_4]\}^{-1}, \tag{8}$$

where α is the maximum specific rate of cell division, θ_i (i=1-4), Φ, and Γ represent the dissimilar contribution of different cells to the regulators production.

For lymphopoiesis model, as both mature lymphocytes and their precursors are radiosensitive [4], each compartment in this system is assumed to be perturbed by radiation [15], and the corresponding equations are:

$$\frac{dx_1}{dt}=Bx_1-\gamma x_1-\frac{N}{D_c}x_1, \tag{9}$$

$$\frac{dx_2}{dt}=\gamma x_1-\delta x_2-\frac{N}{D_2}x_2, \tag{10}$$

$$\frac{dx_3}{dt}=\delta x_2-\kappa x_3-\frac{N}{D_3}x_3, \tag{11}$$

$$\frac{dx_{wd1}}{dt}=\left(\frac{N}{D_c}-\frac{N}{D_1}\right)x_1+Bx_{wd1}-\gamma x_{wd1}-\nu_c x_{wd1}, \tag{12}$$

$$\frac{dx_{di}}{dt}=\frac{N}{D_i}\frac{1}{1+\rho_i}x_i-\nu_1 x_{di}, \tag{13}$$

$$\frac{dx_{hdi}}{dt}=\frac{N}{D_i}\frac{\rho_i}{1+\rho_i}x_i-\nu_2 x_{hdi}, \tag{14}$$

where i=1-3 in eqns (13, 14), and the production rate of X_1 cells

$$B=\alpha\{1+\beta[\theta_1(x_1+x_{wd1}+\Phi x_{d1}+\Gamma x_{hd1})+\theta_2(x_2+\Phi x_{d2}+\Gamma x_{hd2})+\theta_3(x_3+\Phi x_{d3}+\Gamma x_{hd3})]\}^{-1}. \tag{15}$$

The above formalism shows that the granulopoiesis system under chronic radiation can be modeled by kinetics of 7 groups of cells, and the lymphopoiesis system by 10 groups of cells [13, 14]. The control parameters and radiosensitivity parameters are species dependent, which have been estimated based on the information of conventional hematological and radiobiological studies with different species [13, 14].

3 Results and Discussions

We first show the modeling investigation of chronic radiation effects on beagle dogs and our previous attempt to extrapolate the model to humans. The experimental data of chronic radiation effects are retrieved from reports of researchers at Argonne National Laboratory, which documented granulocyte concentrations of beagle dogs over long-term chronic gamma-rays irradiation at various levels [17]. From these data, it is noteworthy that, if the dose-rates are low enough, the granulocytes in the peripheral blood is stabilized at a lower level, and the subjects can survive very high accumulate doses for many years. At levels of 3.0, 7.5, and 18.8 mGy d^{-1} of total body irradiation, the granulocytes concentrations in dogs are reduced to about 65%, 51%, and 41%, respectively, of the normal value [10]. Their experiment also indentified a threshold of dose-rate between 37.5 and 75.0 mGy d^{-1}, beyond which the hematopoietic system may fail [17].

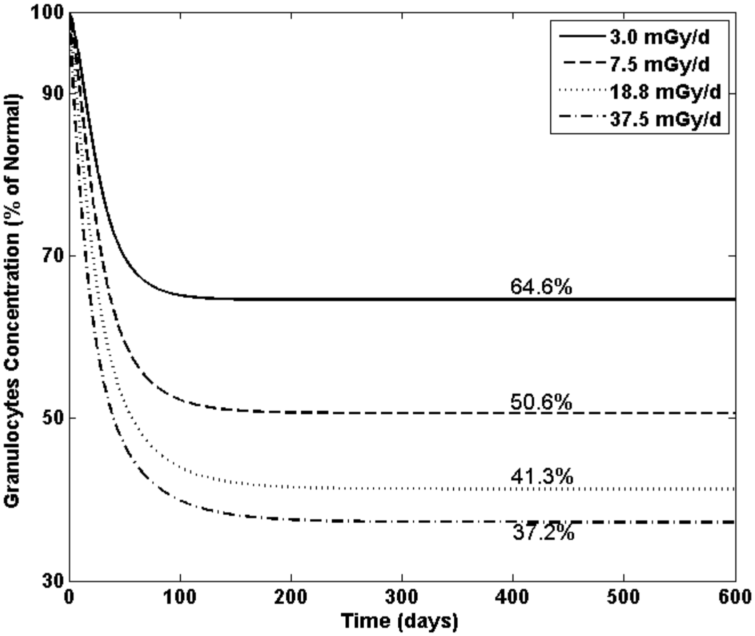


Fig. 1. Modeling the suppression of granulocyte level of beagle dogs under various levels of chronic irradiation, using a dose-rate dependent radiosensitivity scheme of canine granulopoiesis model. The simulated stabilized concentrations of granulocytes are all consistent with experimental data (see text).

For acute and high dose-rate protracted irradiation, we used a fixed radio-sensitivity parameter for the dividing cell pool [13]. We found this scheme does not apply to chronic radiation. If a constant radio-sensitivity parameter is used, the model could not regenerate the suppressed levels of granulocytes consistently. Instead, using a dose-rate dependent radio-sensitivity parameters D_c=30.9N+0.0869 for X_1 cells, and a smaller rate of X_1 cell division α (0.8→0.6) (i.e., assuming a suppressed proliferation capability under chronic radiation), with all other parameters kept the same as in the acute model, the chronic model can accurately simulate all three suppressed levels of granulocytes in beagle dogs at different dose-rates (Figure 1). With these assumptions, the level of depressed granulocyte at a rate of 37.5 mGy d^{-1} is predicted to be about 37% to normal, which is also consistent to the experimental report [17].

The above D_c-N relationship indicates the compartment of the intact X_1 cells acquire certain radio-resistance with the increase of the daily dose-rate. This relation was found by fitting the model to experimental data and, interestingly, happens to be well correlated to the daily dose-rate threshold beyond which the chronic irradiation onto the dogs is life threatening [17]. According to this relationship, the highest dose-rate compatible to the model is about 42.5 mGy d^{-1}. For a dose-rate higher than this threshold, the model becomes unphysical and cannot generate meaningful result.

In a previous study [13], as no hematopoietic data for human could be found in literature for chronic radiation, a similar treatment was applied to human granulopoiesis models, i.e., the X_1 cell division rate α was scaled from 0.6 to 0.5, and a same D_c-N relationship as for canines was applied. We modeled the granulopoietic effects of astronauts under chronic dose-rate 1.5 mSv d^{-1}, which has been predicted for the GCR near solar minimum [5]. Though the level of granulocytes in blood is just slightly depressed (to a level around 90%), according to our model, there will be persistent presence of weakly damaged X_1 cells in bone marrow [13].

Recently a series of results were published for the hematological measurements of the Techa River residents chronically exposed to IR in 1950-1956 [1, 2, and references therein]. These works show distinguished dose-rate effects on the inhibition of different blood cell lines, as well as the deterministic characteristics of the radiation response over a large cohort of human population (Figure 2). Modeling investigation has been conducted with an approach of time delayed differential equations [2]. However, it requires a large number of parameters, such as the steady-state number of hematopoietic stem cells, parameters of feedback regulating function of different cell lines, which are not currently well characterized. Particularly, the model relies upon the contribution of the accumulative dose as well as the dose rates. The animal experiments previously conducted clearly demonstrate the suppressed level of hematopoietic cells are only dose-rate dependent, which could be maintained for many years as long as the dose rates are not life threatening [17]. The results of our previous investigations indicate Smirnova's scheme of hematopoietic models appear to be more appropriate to describe this feature.

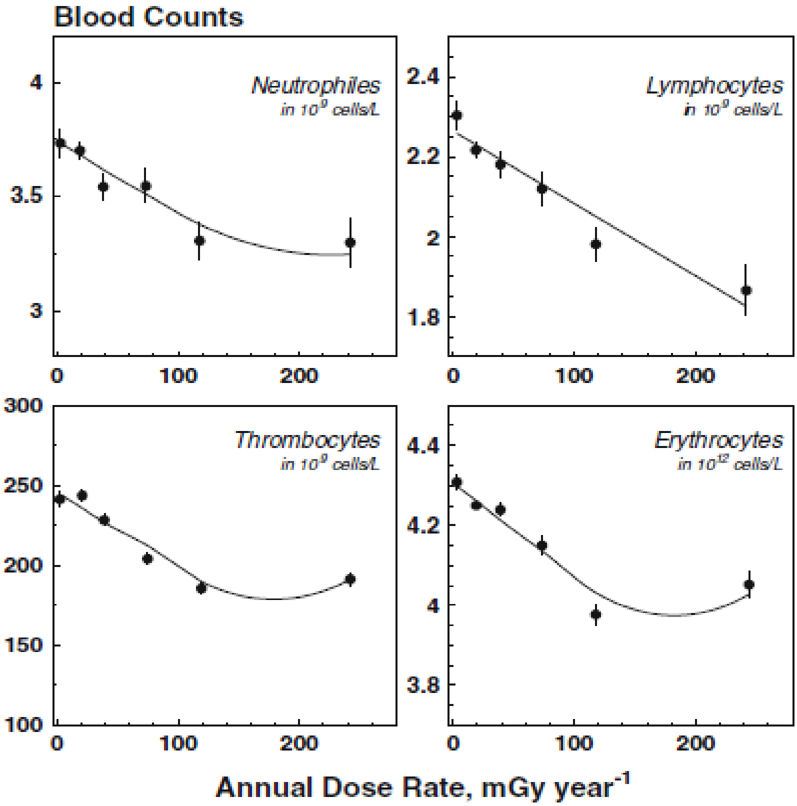


Fig. 2. Data of blood counts for four blood lines, obtained during the years of maximal exposure to radiation (1950-1956) for inhabitants of the Techa River. The figure is taken from [2] with the publisher's permission. The counts of cohorts with no radiation are used as normal values in this study.

To model the depressed levels of lymphocyte counts of the Techa River due to chronic radiation, the chronic lymphopoiesis model assumes a suppressed division rate of X_1 cell (α=0.5 instead of 0.6), and the same D_c-N relationship as for canines, with other control parameters and radiobiological parameters the same as the acute human model previously applied to simulate the empirical data of accidental patients [14]. This treatment is guided from the previous investigation of the canine model for chronic radiation [13]. However, the modeled results are found not consistent with the recorded data. The simulated concentrations of lymphocyte in peripheral blood at all dose-rates are significant higher than the recorded data (Figure 3). This disagreement could be due to several reasons, which include deficiencies in our model, errors in dose reconstruction methods, the possibility of doses from internal emitters including alpha particles, and possible differences in radiation sensitivity between cohorts

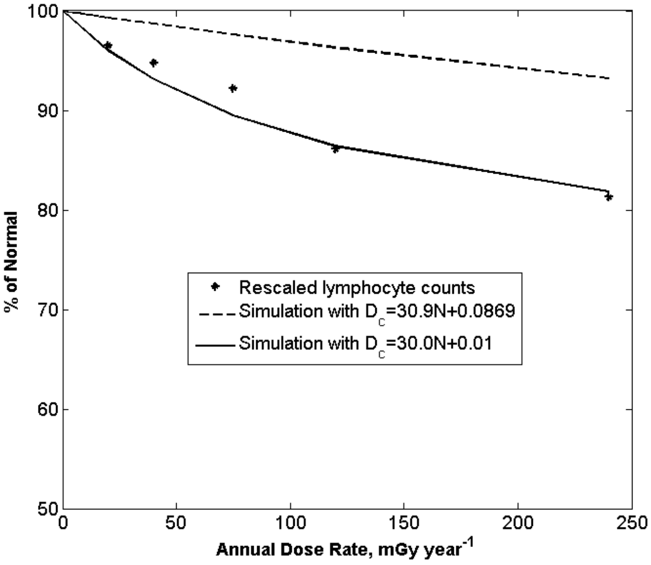


Fig. 3. Rescaled lymphocyte counts of inhabitants of the Techa River subdued by chronic radiation, and model simulations of different schemes. The normal value is assumed 2.3 ×10^9 cells/L.

studied in the report and our estimation. We are unable to account for most of these possibilities, but seek to adjust the scheme used for canine model to improve the model agreement with the radiation response of human hematopoietic system. Though the canine model and human model share a same general structure, their parameters are much different from each other [13].

We thus tried to modify the D_c-N relationship to see whether a better fitting between the model and recorded data could be achieved. As the results indicate the human precursor cells are apparently more sensitive to chronic radiation than canines (Figure 3), the value of D_c for human at a dose rate N should be smaller than the corresponding value for canine. A modified D_c-N relationship is found to provide a better fitting between the model and recorded data at all dose rates (Figure 3).

As the granulopoiesis model shares a same coarse-grained hematopoietic structure with the lymphopoiesis model, and the X_1 cells in both models refer to the same precursor cells cohort, it is vital to check whether this modified Dc-N relationship also applies to the granulopoiesis system. The Techa River data provides the neutrophile counts for groups under different dose rates (Figure 2) [2]. As neutrophiles are the main components (up to 99%) of granulocytes, their kinetics under chronic radiation should follow a same scheme as the granulopoiesis model. We found, if the canine Dc-N relationship is applied, the simulated inhibition neutrophiles due to

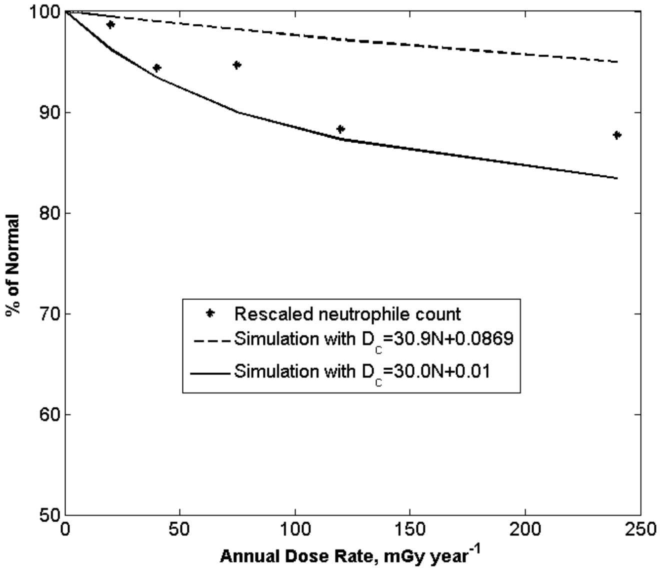


Fig. 4. Rescaled neutrophile counts of inhabitants of the Techa River subdued by chronic radiation, and model simulations of different schemes. The normal value is assumed 3.75 ×10^9 cells/L.

chronic radiation is also significantly underestimated (Figure 4). On the other hand, the modified Dc-N relationship gives a much better fitting between model and recorded data (Figure 4). This demonstrates a unified dose-rate dependent radiosensitivity parameter applies to both of the lymphopoiesis and granulopoiesis systems for humans under chronic radiation, and confirms that, barring possible dosimetry errors in the Techa river study, the human hematopoietic precursor cells are more sensitive to chronic radiation than the cell compartment in the previously well-characterized canine model.

4 Conclusions and Future Work

A unified dose-rate dependent radiosensitivity parameter is found to be applicable to both lymphopoiesis and granulopoiesis systems for human. This is not unexpected as both models follow a same coarse-grained scheme of hematopoietic structure, and particularly, the mature lymphocytes and granulocytes in peripheral blood are developed from a same precursor cell cohort [3]. Moreover, this modeling investigation on the Techa River hematological data indicates the previous assumption of the radiosensitivity parameter for human hematopoietic stem cells need to be corrected [13]. The Techa riverside area is known as the only region in the world where a significant number of the population suffered from chronic radiation for a long time period.

Though a historical tragedy, the collected data and consequent analysis can be of great value to the study of the mechanisms involved in the development of chronic radiation effects in humans, which is essential to radiation protection in future human space exploration.

In addition to the lymphocyte and granulocyte data, the Techa River hematological data also contains platelet and erythrocyte counts [2]. It will be of interest to investigate whether the hematopoietic scheme of this study also applies to these cellular lines. Based on our previous studies, this scheme appears to have the potential to build up a unified model to characterize mammalian hematopoietic response to various scenarios of radiation, which has been pursued by many researchers for several decades [4, 9].

Acknowledgement. This study was supported by the NASA Space Radiation Risk Assessment Project.

References

1. Akleyev, A.V., Akushevich, I., Dimov, G.P., Veremeyeva, G.A., Varfolomeyeva, T.A., Ukraintseva, S.V., Yashin, A.I.: Early hematopoiesis inhibition under chronic radiation exposure in humans. Radiat. Environ. Biophys. 49(2), 281–291 (2010)
2. Akushevich, I.V., Veremeyeva, G.A., Dimov, G.P., Ukraintseva, S.V., Arbeev, K.G., Akleyev, A.V., Yashin, A.I.: Modeling hematopoietic system response caused by chronic exposure to ionizing radiation. Radiat. Environ. Biophys. 50, 299–311 (2011)
3. Beutler, E., Coller, B.S., Lichtman, M.A., Kipps, T.J., Seligsohn, U. (eds.): Williams Hematology, 6th edn. McGraw-Hill, New York (2001)
4. Bond, V.P., Fliedner, T.M., Archambeau, J.O.: Mammalian radiation lethality. Academic Press, New York (1965)
5. Cucinotta, F.A., Kim, M.Y., Ren, L.: Evaluating Shielding Effectiveness for Reducing Space Radiation Cancer Risks. Radiat. Meas. 41, 1173–1185 (2006)
6. Cucinotta, F.A., Durante, M.: Cancer risk from exposure to galactic cosmic rays: implications for space exploration by human beings. Lancet Oncol. 7, 431–435 (2006)
7. Cucinotta, F.A., Chappell, L.J.: Updates to Radiation Risk Limits for Astronauts: Risks for Never-smokers. Radiat. Res. 176, 102–114 (2011)
8. Durante, M., Cucinotta, F.A.: Heavy ion carcinogenesis and human space exploration. Nat. Rev. Cancer 8, 465–472 (2008)
9. Fliedner, T.M., Graessle, D., Meineke, V., Dorr, H.: Pathophysiological principles underlying the blood cell concentration responses used to assess the severity of effect after accidental whole-body radiation exposure: an essential basis for an evidence-based clinical triage. Exp. Hematol. 35(4 suppl. 1), 8–16 (2007)
10. Fliedner, T.M., Tibken, B., Hofer, E.P., Paul, W.: Stem cell responses after radiation exposure: a key to the evaluation and prediction of its effects. Health Phys. 70, 787–797 (1996)
11. Fliedner, T.M., Graessle, D., Paulsen, C., Reimers, K.: Structure and function of bone marrow hemopoiesis: mechanisms of response to ionizing radiation exposure. Cancer Biother. Radiopharm. 17, 405–426 (2002)
12. Hu, S., Cucinotta, F.A.: A cell kinetic model of granulopoiesis under radiation exposure: Extension from rodents to canines and humans. Radiat. Prot. Dosimetry 143, 207–213 (2011)

13. Hu, S., Cucinotta, F.A.: Characterization of the radiation-damaged precursor cells in bone marrow based on modeling of the peripheral blood granulocytes response. Health Phys. 101, 67–78 (2011)
14. Hu, S., Smirnova, O.A., Cucinotta, F.A.: A biomathematical model of lymphopoiesis following severe radiation accidents — potential use for dose assessment. Health Phys. 102, 425–436 (2012)
15. Kovalev, E.E., Smirnova, O.A.: Estimation of radiation risk based on the concept of individual variability of radiosensitivity. AFRRI Contract Report 96-1. Armed Forces Radiobiology Research Institute, Bethesda (1996)
16. National Council on Radiation Protection and Measurements. Information needed to make radiation protection recommendations for space missions beyond low-earth orbit. NCRP, Bethesda, MD, Report No. 153 (2006)
17. Seed, T.M., Fritz, T.E., Tolle, D.V., Jackson III, W.E.: Hematopoietic response under protracted exposure to low daily dose gamma irradiation. Adv. Space Res. 30, 945–955 (2002)
18. Smirnova, O.A.: Environmental radiation effects on mammals: a dynamical modeling approach. Springer, New York (2011)
19. Till, J.E., McCulloch, E.A.: A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiat. Res. 14, 213–222 (1961)
20. Zukhbaya, T.M., Smirnova, O.A.: An experimental and mathematical analysis of lymphopoiesis dynamics under continuous irradiation. Health Phys. 61, 87–95 (1991)

Feature-Based Medical Image Registration Using a Fuzzy Clustering Segmentation Approach

Hassan Mahmoud, Francesco Masulli*, and Stefano Rovetta

Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, Università di Genova, and CNISM Genova Research Unit, Via Dodecaneso 35, I-16146 Genova, Italy
* Center for Biotechnology, Temple University, Philadelphia, USA
{hassan.mahmoud,francesco.masulli,stefano.rovetta}@unige.it

Abstract. This paper presents an approach to medical image registration using a segmentation step based on Fuzzy C-Means (*FCM*) clustering and the Scale Invariant Feature Transform (*SIFT*) for matching keypoints in segmented regions. To obtain robust segmentation, FCM is applied on feature vectors composed by local information invariant to image scaling and rotation, and to change in illumination. *SIFT* is then applied to corresponding regions in reference and target images, after the application of an $alpha$-cut. The proposed registration method is more robust to noise artifacts than standard *SIFT*. The paper shows also a method for *FCM* clustering speeding-up based on a dynamic pyramid approach using low resolution images of increasing size.

1 Introduction

Image registration [13] is the process of aligning images so that corresponding features can easily be related. In medical imaging it allows us to extract complementary information from different modalities, and to compare accurately images from the same modality [4,10]. Recently, registration has been also applied in image guided surgery interventions, and in serial imaging analysis for the study of diseases progression.

To achieve image registration, the computer rotates, scales and translates one image (*target image*) to match another image (*reference image*). Methods to perform the registration can be categorized as feature-based, intensity-based, and gradient-based, although hybrid approaches are possible [13]. In feature-based methods the registration is based on the correspondence of a small set of salient points, landmarks, or on alignment of segmented binary structures in images being registered (e.g., lines, curves or points matching). These methods are relatively fast, but they are lacking in robustness of feature extraction and accuracy of feature matching. Furthermore, extracted features need to be invariant to image deformations. To this aim, because of the noise entailed in medical images, some preprocessing steps are usually applied to enhance feature appearance using image gradients and gamma corrections. In particular, the accuracy of the registration result depends on the quality of the previous region segmentation procedure.

Medical image segmentation methods are usually based on gray level features (e.g., histograms, edges, regions), texture features (e.g., first or higher order statistics, spectral methods) correspondence, or also on model based or atlas based techniques [13,14].

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 37–47, 2013.

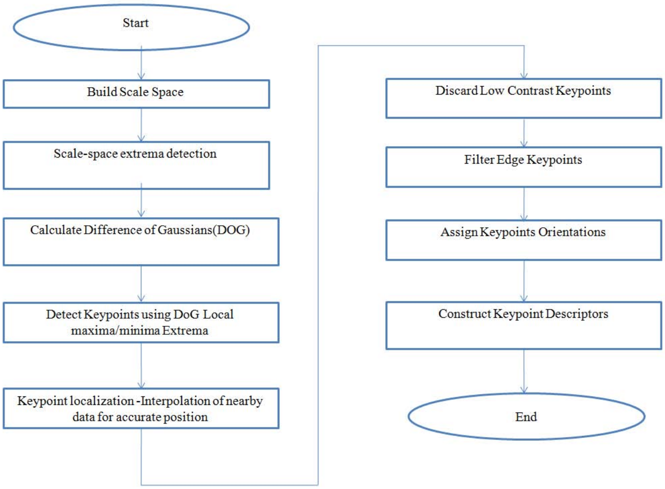

Start
Build Scale Space
Scale-space extrema detection
Calculate Difference of Gaussians(DOG)
Detect Keypoints using DoG Local maxima/minima Extrema
Keypoint localization -Interpolation of nearby data for accurate position
Discard Low Contrast Keypoints
Filter Edge Keypoints
Assign Keypoints Orientations
Construct Keypoint Descriptors
End

Fig. 1. *SIFT* feature detection technique

Recently, artificial neural network methods and clustering techniques have been successfully applied [11,9].

In this paper we apply fuzzy clustering to automatically detect robust candidate regions for a registration method based on the Scale Invariant Feature Transform (*SIFT*) [6] that is a popular feature-based image registration method matching points using a similarity measure.

This paper is organized as follows: The Scale Invariant Feature Transform and Fuzzy C-Means clustering algorithm are presented in Sect.s 2 and 3; Sect. 4 presents the proposed *FCM–SIFT* registration framework; results and discussion are in Sect. 5; Sect. 6 contains the conclusions.

2 Scale Invariant Feature Transform

In Scale-Invariant Feature Transform (*SIFT*) [6], robust and salient reference points (keypoints) of objects are extracted from the reference image and from a target image to be co-registered respect to it. Fig. 1 shows our implementation of *SIFT* to detect keypoints from corresponding segmented regions in the reference and target images. The main steps are:

1. *Scale-space extrema detection*. In this step we search over all scales and image locations by using a Difference-of-Gaussian function (DoG) to identify potential interest points that are invariant to scale and orientation. We compare each pixel in the DoG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales. If the pixel value is the

maximum or minimum among all compared pixels, it is selected as a candidate key-point. Specifically, a DoG image $D(x, y, \sigma)$ is given by:

$$D(x, y, \sigma) = L(x, y, k_i\sigma) - L(x, y, k_j\sigma), \tag{1}$$

where $L(x, y, k\sigma)$ is the convolution of the original image $I(x, y)$ with the Gaussian kernel

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

at scale $k\sigma$, i.e.,

$$L(x, y, k\sigma) = G(x, y, k\sigma) * I(x, y) \tag{3}$$

Note that this first step produces many keypoints.

2. *Keypoint localization*. Keypoints are selected using measures of their stability, using nearby data, scale, and ratio of principal curvatures. This information allows points to be rejected that have low contrast and are sensitive to noise or poorly localized along an edge.
3. *Orientation assignment*. Orientations are then assigned on the basis of local image gradient directions to each key-point for rotation invariance. *SIFT* operates on data transformed to the assigned orientation, scale, and location for each feature, providing invariance to these transformations. For an image sample $L(x, y)$ at scale σ, the gradient magnitude, $m(x, y)$, and orientation $\Theta(x, y)$, are pre-computed using pixel differences:

$$m(x, y) = \sqrt{[(L(x+1, y, \sigma) - L(x-1, y, \sigma)]^2 + [(L(x, y+1, \sigma) - L(x, y-1, \sigma)]^2} \tag{4}$$

and

$$\Theta(x, y) = tan^{-1}\left(\frac{L(x+1, y, \sigma) - L(x-1, y, \sigma)}{L(x, y+1, \sigma) - L(x, y-1, \sigma)}\right) \tag{5}$$

where $L(x, y, \sigma)$m is the Gaussian smoothed image.

After applying *SIFT* on target and reference images we obtain the set of salient feature points. It is worth noting that the quality of *SIFT* results, as for other feature-based methods for registration, is strongly affected by the quality of the previous region segmentation procedure.

3 Fuzzy C-Means Algorithm

The determination of consistent clusters, i.e., matched segments/regions in the reference and target images is a main step in *SIFT*. In [6], clustering is performed by the generalized Hough transform. In the approach we propose in this paper clustering is obtained using the Fuzzy C-Means [1] (*FCM*) clustering algorithm.

The *FCM* algorithm is aimed to the minimization of the following functional:

$$J_m(\mathbf{U}, Y) \equiv \sum_{i=1}^{n} \sum_{k=1}^{c} (u_{ik})^m E_k(x_i) \tag{6}$$

where: $X = \{x_1, x_2, \ldots, x_n\}$ is a data set containing n unlabeled sample points; $Y = \{y_1, y_2, \ldots, y_c\}$ is the set of the centers of clusters; $\mathbf{U} = [u_{ik}]$ is the $c \times n$ fuzzy c-partition matrix, containing the membership values of all samples to all $m \in (1, \infty)$ is the fuzziness control parameter; $E_k(x_i)$ is a dissimilarity measure (distance or cost) between data point x_i and the center y_k of a specific cluster k. We use the Euclidean distance $E_k(x_i) = \|x_i - y_k\|^2$ as the dissimilarity measure.

The clustering problem can be formulated as the minimization of J_m with respect to Y, under the normalization constraint $\sum_{k=1}^{c} u_{ik} = 1$.

The necessary conditions for minimization of J_m are then:

$$y_k = \frac{\sum_{i=1}^{n} (u_{ik})^m x_i}{\sum_{i=1}^{n} (u_{ik})^m} \qquad \text{for all } k, \tag{7}$$

$$u_{ik} = \begin{cases} \left(\sum_{l=1}^{c} \frac{E_k(x_i)}{E_l(x_i)} \right)^{\frac{2}{1-m}} & \text{if } E_k(x_i) > 0 \, \forall k, i; \\ 1 \quad \text{if } E_k(x_i) = 0 \quad \text{and} \quad u_{il} = 0 \ \forall \, l \neq k \end{cases} \tag{8}$$

The Fuzzy C-Means algorithm starts with a random initialization of the fuzzy c-partition matrix $\mathbf{U}$ (or of the centroids y_k) and then implements a Picard iteration of Eq.s 7 and 8 until convergence (defined, e.g., as when the change of centroids is smaller than an assigned threshold).

Note that if one chooses $m = 1$ the Fuzzy C-Means functional J_m (Eq. (6)) reduces to the expectation of the K-Means (*KM*) global error $< E > \equiv \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik} E_k(x_i)$, and the *FCM* becomes the crisp *KM* algorithm [15,5,3].

4 Fuzzy C-Means Based Scale Invariant Feature Transform

As already stated, in our proposed approach for image registration *SIFT* operates on the matched segments (clusters) obtained from *FCM*. Starting from those segments, *SIFT* extracts the matching keypoints in both reference and target images and obtains the registration parameters able to recover their correspondence.

In order to find robust and reliable clusters, *FCM* must be performed in a feature space with features invariant to image scaling and rotation. In our approach, for each pixel we consider intensity value, spatial location, and neighborhood average intensity and deviation from the eight surrounding pixels. These features are well localized in both the spatial and frequency domains (reducing the probability of disruption by occlusion, clutter, or noise), are invariant to image scaling and rotation, and are partially invariant to moderate changes in illumination and 3D camera viewpoint [2,12,7].

After clustering the two images, we select the minimal volume (fuzzy cardinality) cluster, corresponding to a region in each image, and then we apply an α-cut, where α is a threshold selected in the interval $[0, 1]$ that selects pixels with high membership to cluster. This approach minimizes the search space for finding the correspondence of keypoints, as it selects the most reliable pixels of the region. Note that, if $\alpha = 0$ we get the whole pixels of the segmented region, while if $\alpha = 1$ we select pixels having stronger membership to the cluster only, but this sub-set could be empty or could lose

Table 1. The *FCM–SIFT* registration framework

1. *Image Preprocessing*: Adjust threshold gray level enhancement in both target and reference images and obtain image pyramid.
2. *Segment Extraction*: Apply *FCM* on target and reference images with c clusters and obtain **U** fuzzy membership matrix.
3. *Segment matching*:
   (a) Calculate cluster volumes from the fuzzy membership matrix **U**.
   (b) Find the minimum volume clusters in both images (which represent the smallest matched region in the two images) and and extract two reliable segments from them withan α-cut.
4. *Feature matching*: Apply *SIFT* on both segments to extract invariant robust feature points.
5. *Registration*:
   (a) Infer spatial correspondence between at least two points of extracted matched feature points in both images.
   (b) Extract registration parameters.
   (c) Apply spatial transformation on target image.

Table 2. Experiments. For each experiment we report: horizontal translation (Tx), vertical translation (Ty), rotation degree (R), and scaling factor (S).

Experiment	Tx	Ty	R	S
T1	20	20	252	1.0
T2	-30	-30	0	.6
T3	40	0	324	1.3
T4	0	-10	144	.8
T5	-20	-20	120	.7

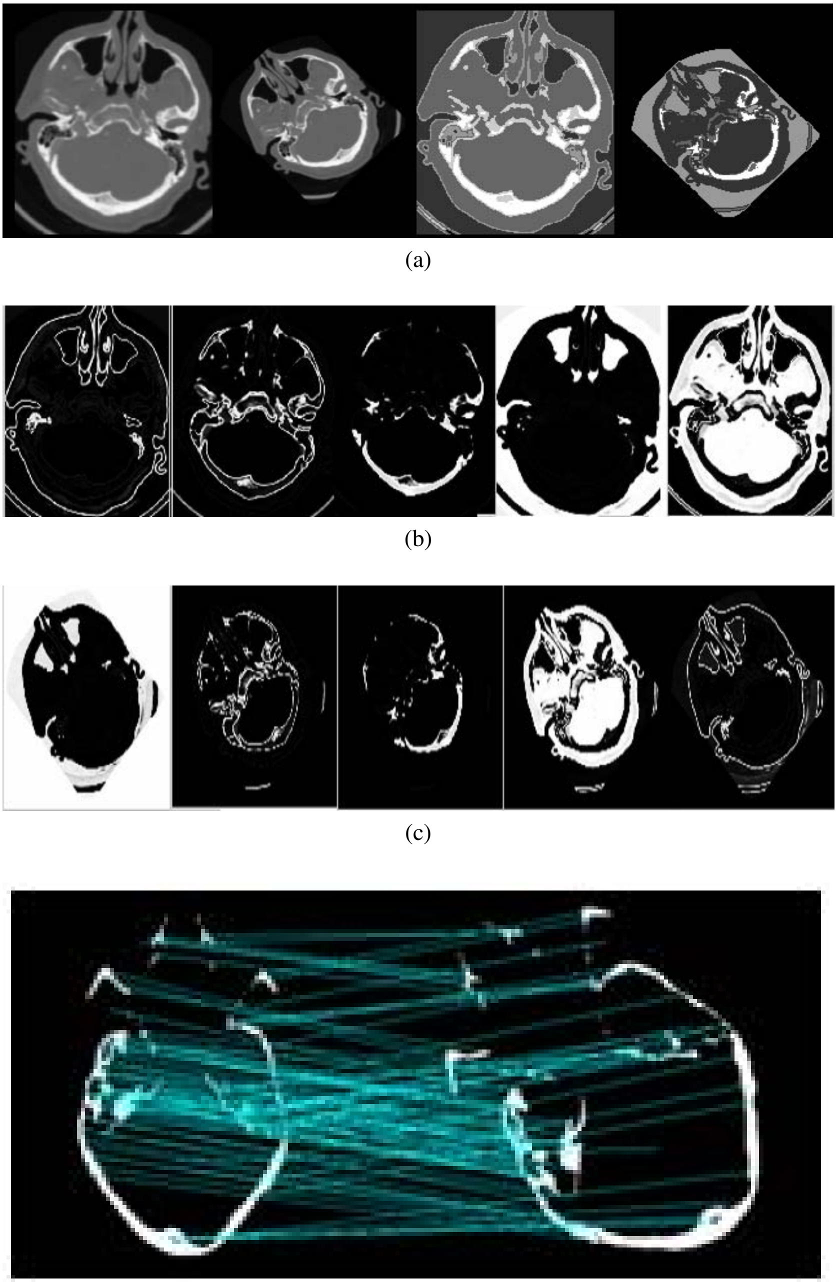
good keypoint candidates. Therefore, we choose the highest value of α that select a region with a size corresponding to an assigned percentage of the full image.

The steps of the proposed *FCM–SIFT* registration framework are described in Tab. 1. At the end of the *FCM–SIFT* registration, we obtain the registration parameters between target and reference images (namely, horizontal translation Tx, vertical translation Ty, rotation angle R, and scaling factor S).

5 Experimental Results and Discussion

We validated our *FCM–SIFT* registration framework on a sample set of Computer Tomography (CT) images of the head. The software was developed in Matlab R2009b under Windows 7 32 bit. The computation time ($Time$) was evaluated on a laptop with 2.00 GHz dual-core processor and 3.25 GB of RAM. As usual, time is given as a rough indication only, with the additional caveat that Matlab is inefficient in specific operations, for instance loops.

The target images where obtained by transforming the original image with a combination of translation, rotation, and anisotropic scaling.



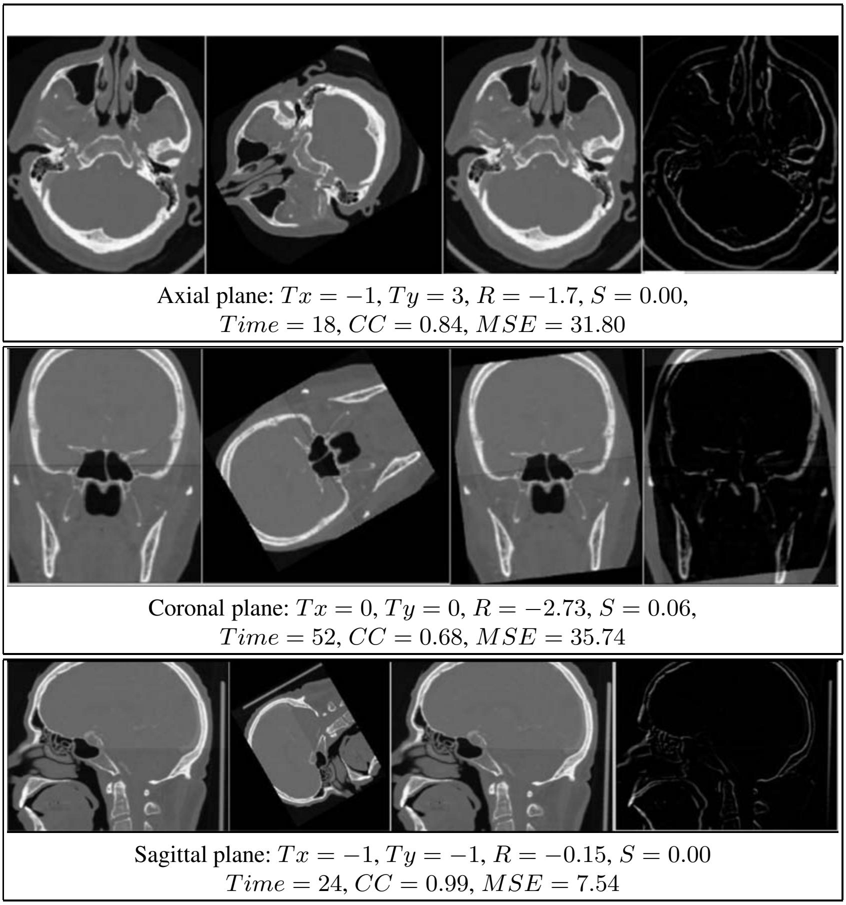
(a)

(b)

(c)

(d)

Fig. 2. Axial CT head plane. (a) From left to right: Reference image, target image, segmented reference image, and segmented target image. (b) The five clusters obtained from the reference image. (c) The five clusters obtained from the target image. (d) *SIFT* matching using the minimal volume regions, after α-cut.



Axial plane: $Tx = -1, Ty = 3$, $R = -1.7$, $S = 0.00$,
$Time = 18$, $CC = 0.84$, $MSE = 31.80$

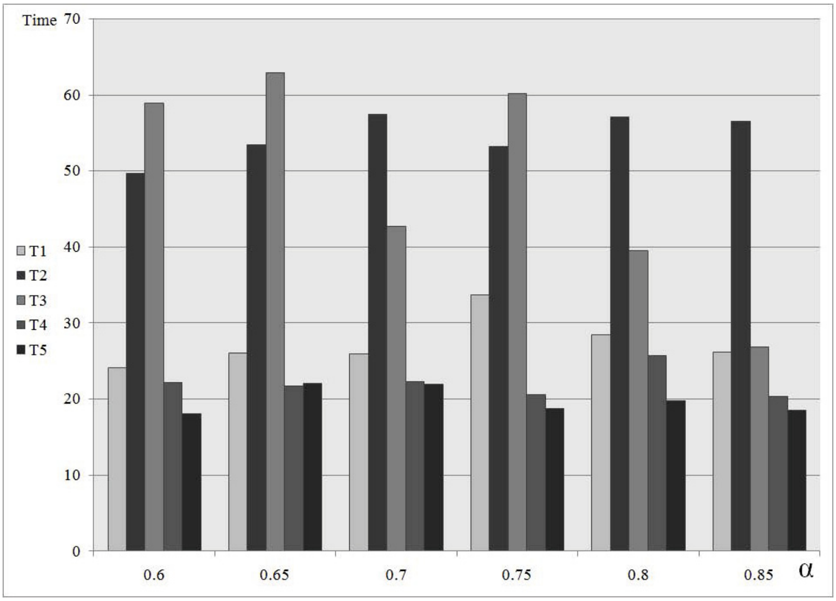
Coronal plane: $Tx = 0, Ty = 0$, $R = -2.73$, $S = 0.06$,
$Time = 52$, $CC = 0.68$, $MSE = 35.74$

Sagittal plane: $Tx = -1, Ty = -1$, $R = -0.15$, $S = 0.00$
$Time = 24$, $CC = 0.99$, $MSE = 7.54$

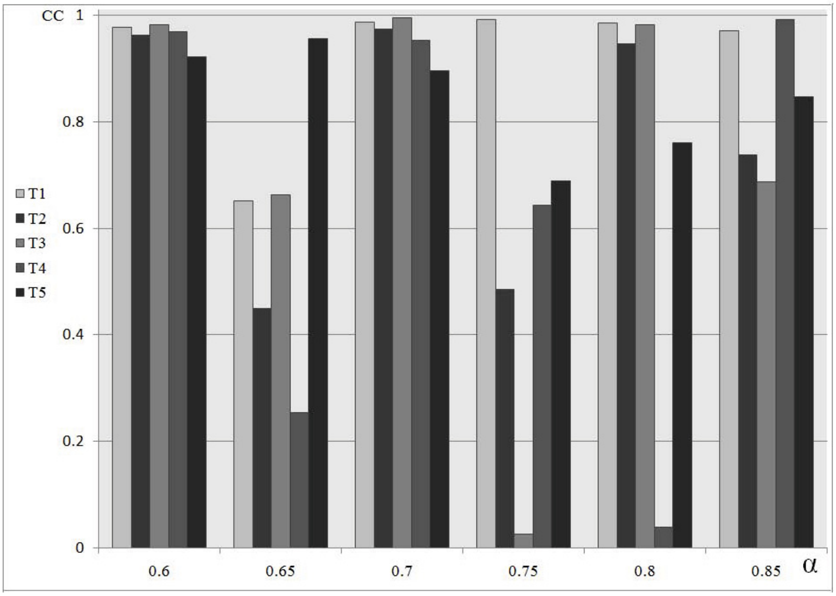
Fig. 3. *FCM–SIFT* registration results with CT of head on axial, sagittal, and coronal planes. From left to right: reference, moving, registered, and error images.

Fig. 2 shows the segment extraction on axial head CT slice on reference and target images using *FCM* clustering. In Fig. 2a, from left to right, there are the reference image, the target image, the segmented reference image, and the segmented target image. *FCM* is performed using a number of clusters $c = 5$ estimated on the basis of a-priori experimental knowledge and a fuzzyfication parameter $m = 2$. Fig.s 2b and 2c show the five clusters obtained from reference and target images. After clustering, we identified in the two images the segments with minimal volumes to be matched. Then we selected the regions with points with highest memberships by applying the α-cut thresholding. Finally, we applied the *SIFT* on this pair of sub-regions. Fig. 2d shows *SIFT* matching of salient keypoints of the selected regions.

Fig. 3 illustrates some *FCM–SIFT* registration results with CT on axial, sagittal, and coronal planes. For each projection, we report, from left to right, the reference and target images, the registered image, and the error image defined as the difference between the

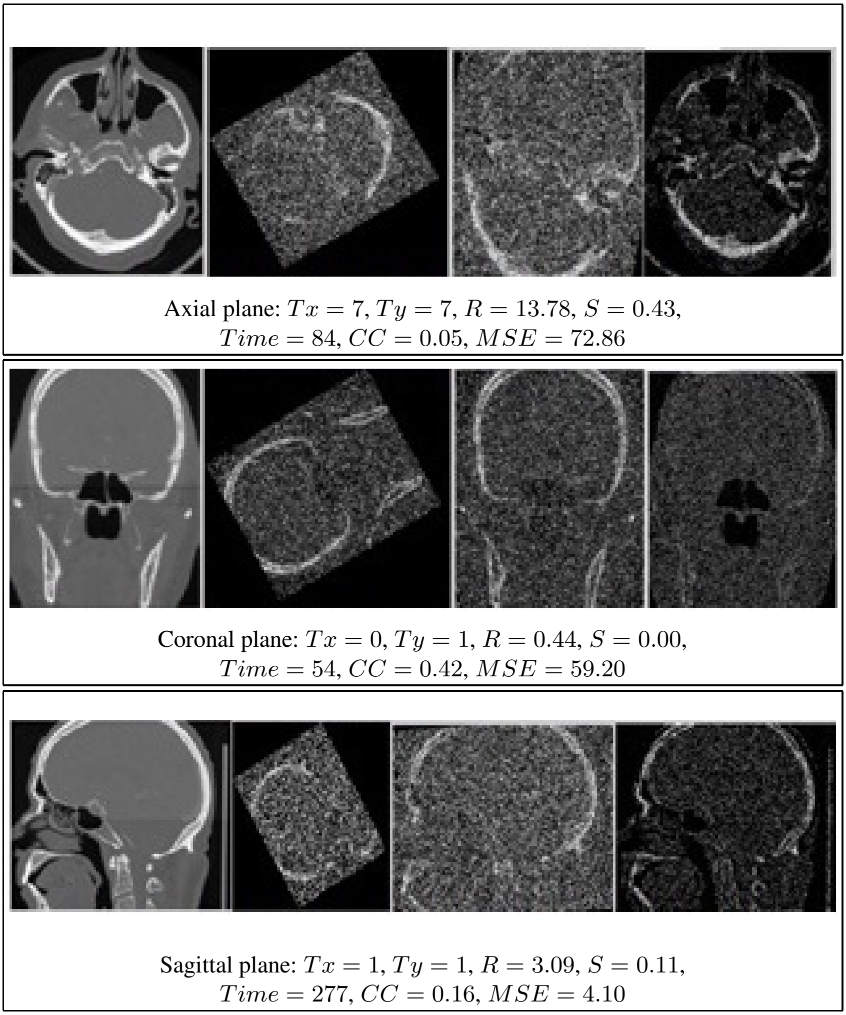


(a)



(b)

Fig. 4. Experiments T1–T5: Running time in seconds (a) and cross correlation between registered and reference images (b) v.s. threshold α



Axial plane: $Tx = 7, Ty = 7$, $R = 13.78$, $S = 0.43$,
$Time = 84$, $CC = 0.05$, $MSE = 72.86$

Coronal plane: $Tx = 0, Ty = 1$, $R = 0.44$, $S = 0.00$,
$Time = 54$, $CC = 0.42$, $MSE = 59.20$

Sagittal plane: $Tx = 1, Ty = 1$, $R = 3.09$, $S = 0.11$,
$Time = 277$, $CC = 0.16$, $MSE = 4.10$

Fig. 5. *FCM–SIFT* registration results with CT of head on axial, sagittal, and coronal planes. From left to right: reference, target with salt and pepper noise ($\nu = 0.5$), registered, and error images.

reference and the registered image. We show also the values of the cross correlation (CC), the root mean square error (MSE), the computation time ($Time$), the value of threshold of the α-cut, the horizontal translation (Tx), the vertical translation (Ty), the rotation degree (R), and the scaling factor (S).

In Fig. 4, we report the results of five experiments (T1–T5), using the axial CT of head illustrated in Fig. 2. The target image is obtained by applying the transformations shown in Tab. 2. Fig. 4a shows the dependence of running time of the *FCM–SIFT* technique on the value of α, that is the threshold of the α-cut. Fig. 4b, in turn, shows the dependence of the cross correlation (CC) between reference and registered images obtained using the *FCM–SIFT* technique on the value of α-cut.

To study the noise robustness of our approach to segmentation, we applied the *FCM–SIFT* technique on noisy orthogonal slices in axial, sagittal, and coronal planes, by adding salt and pepper noise (impulse noise) to the target image. This kind of noise is typically observed on advanced medical imaging equipments such as CT, MRI (Magnetic Resonance Imaging) and PET (Positron Emission Tomography). It appears as randomly occurring sparse light and dark disturbances in the image (white and black pixels). Typical sources include flecks of dust inside the camera, or, with digital cameras, faulty CCD (Charge-Coupled Device) image sensors elements.

We compared the results of registration in the presence of salt and pepper noise using our proposed approach *FCM–SIFT*, a modified version of our approach using *KM* instead of *FCM* (*KM–SIFT*), and standard *SIFT*. In our experiments we notice that the *breakdown point*[1] for standard *SIFT* and *KM–SIFT* is a noise density value $\nu = .4$, as those methods cannot detect the correspondence between reference and target images in the presence of a noise density higher than this value, while the breakdown point of *FCM–SIFT* is $\nu = .59$. Fig. 5 shows the of *FCM–SIFT* registration results in the presence of salt and pepper noise ($\nu = .50$). To increase the breakdown point of *FCM–SIFT* we have to increase α; but this is possible until a critical value where the registration results degrade as we may cut off relevant features.

As shown in Fig. 4a, the average registration time in *FCM–SIFT* is about 30 seconds, while the average registration time in standard *SIFT* is about 4 seconds. For speeding up the clustering phase, we have experimented also a dynamic pyramid approach for clustering, applying *FCM* on low resolution images of increasing size. This pyramidal approach to *FCM* can reduce the average registration time to about 10 seconds.

For speed up the clustering phase, we use a dynamic pyramid approach that allow us to operate on reduced images instead of original images, thus reducing clustering and registration times. Then we reconstruct the pyramid and register the original images after calculating the registration parameters from reduced resampled images obtained from scale resolution pyramid.

6 Conclusions

Medical image registration procedures allow us to extract complementary information from different modalities, and to accurately compare images from the same modality [4,10].

This paper proposes an approach to medical image registration using a segmentation step segmentation based on Fuzzy C-Means (*FCM*) clustering [1] and Scale Invariant Feature Transform (SIFT) [6] for matching keypoints in segmented regions.

It is worth noting that the quality of *SIFT* results, as for other feature-based methods for registration, is strongly affected by the quality of the previous region segmentation procedure. To obtain robust segmentation, we applied FCM feature vectors including

[1] We use here this term, borrowed from Robust Statistics [8], as the minimum value of noise density that makes the *SIFT* procedure unsuccessfully.

local information invariant to image scaling and rotation, and to change in illumination [2,12,7].

The paper shows also how to reduce the running time of the clustering step following a dynamic pyramid approach applying *FCM* on low resolution images of increasing size. The reported speed-up is about three.

Medical images are often corrupted by noise; in particular, salt and pepper noise is typically observed on advanced medical imaging equipments. The robustness of algorithms with respect to noise is then a major request in medical imaging. From our experimental results, we can conclude that the proposed *FCM–SIFT* registration method is more robust to noise artifacts than standard *SIFT* and a modified version of our approach using *KM* instead of *FCM*.

Acknowledgements. Work partially funded by a grant of the University of Genova. Hassan Mahmoud is a PhD student in Computer Science at DIBRIS, University of Genova.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
2. Cai, W., Chen, S., Zhang, D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recognition 40, 825–838 (2007)
3. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
4. Hallpike, L., Hawkes, D.J.: Medical image registration: an overview. Imaging 14, 455–463 (2002)
5. Lloyd, S.: Least square quantization in PCM's. Bell Telephone Laboratories Paper (1957); also IEEE Trans. Inform. Theory 28, 129–137 (1982)
6. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, pp. 1150–1157 (1999)
7. Foo, J.L., Miyano, G., Lobe, T., Winer, E.: Three-dimensional segmentation of tumors from CT image data using an adaptive fuzzy system. Computers in Biology and Medicine 39, 869–878 (2009)
8. Huber, P.J.: Robust Statistics. Wiley (1981)
9. Ji, Z.-X., Sun, Q.-S., Xia, D.-S.: A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain MR image. Computerized Medical Imaging and Graphics 35, 383–397 (2011)
10. Maintz, J.B.A., Viergever, M.A.: A survey of Medical Image Registration. Medical Image Analysis 2, 1–36 (1998)
11. Masulli, F., Schenone, A.: A fuzzy clustering based segmentation system as support to diagnosis in medical imaging. Artificial Intelligence in Medicine 16, 129–147 (1999)
12. Moreno, A., Takemura, C.M., Colliot, O., Camara, O., Bloch, I.: Using anatomical knowledge expressed as fuzzy constraints to segment the heart in CT images. Pattern Recognition 41, 2525–2540 (2008)
13. Sharma, N., Ray, A.K., Sharma, S., Shukla, K.K., Pradhan, S., Aggarwal, L.M.: Segmentation and classification of medical images using texture-primitive features: Application of BAM-type artificial neural network. J. Med. Phys. 33, 119–126 (2008)
14. Sharma, N., Aggarwal, L.M.: Automated medical image segmentation techniques. J. Med. Phys. 35, 3–14 (2010)
15. Steinhaus, H.: Sur la division des corp materiels en parties. Bulletin de l'Academie Polonaise des Sciences, C1. III IV, 801–804 (1956)

A Novel Approach Based on Joint Optimization of Alignment and Statistical Surface Representation with Wavelet Transform for CBCT Segmentation

Yu-Bing Chang[1,2,4], Peng Yuan[2], Tai-Hong Kuo[3], Zixiang Xiong[1],
Jaime Gateno[2], James J. Xia[2], and Xiaobo Zhou[4,⋆]

[1] Department of Electrical and Computer Engineering, Texas A&M University,
College Station, TX 77841, USA

[2] Department of Oral and Maxillofacial Surgery, The Methodist Hospital Research
Institute, Houston, TX 77030, USA

[3] Department of Mechanical Engineering, National Cheng Kung University, Tainan,
Taiwan

[4] Center for Biotechnology and Informatics, The Methodist Hospital Research
Institute, Houston, TX 77030, USA

XZhou@tmhs.org

Abstract. Cone-beam computed tomography (CBCT) can provide true 3D information of anatomical structures, with advantages of much thinner slice thickness and significantly lowered effective dose of radiation. However, CBCT images are extremely low contract and noisy. It is very difficult to segment thin bones. It usually takes 4-5 hours to manually segment a set of CBCT data. To this end, we developed a novel approach based on the joint optimization of alignment and statistical surface representation with wavelet transform for segmentation of CBCT images. It included two main steps: customized wavelet base initialization (CWBI) and base invariant wavelet active shape model (BIWASM). We validated our approach with others by comparing the surface deviation between segmented shape to the ground truth. The results showed that our approach outperformed the others in accuracy and computing time.

Keywords: Segmentation, Cone-beam CT, Statistical Shape Model.

1 Introduction

Computed tomography (CT) are commonly used for treatment planning of patients with craniomaxillofacial and dentofacial deformities [11]. With the giant leap of technology, CT scans are gradually replaced by cone-beam computed tomography (CBCT) scans. Like conventional CT, CBCT has true 3D information of the anatomical structures. In addition, CBCT slice thickness (0.125mm) is much thinner than CT (0.625mm). CBCT is also a safer imaging modality.

[⋆] Corresponding author.

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 48–56, 2013.

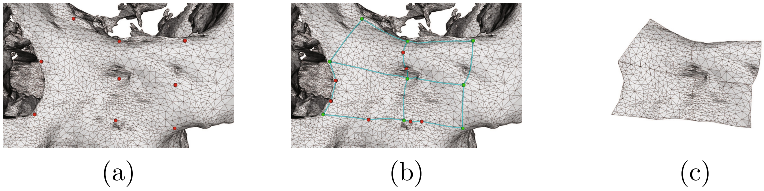


Fig. 1. Shape extraction and patch decomposition. (a) 9 control landmark points. (b) Patch decomposition using shortest paths as boundaries. (c) Extraction of patches and shape.

Its effective dose of for a whole head scan (≈50gSv) is significantly lower than CT's (≈4500gSv) [11], [10]. However, CBCT images are very low contract and noisy [9]. This makes extremely difficult to segment thin bones (i.e., anterior walls of maxillary sinus) from the background without picking up unwanted noises. Therefore, the purpose of this project is to develop a novel approach based on the joint optimization of alignment and statistical surface representation with diffusion wavelet transform for segmentation of CBCT images.

2 Customized Wavelet Distribution Model

2.1 Shape Customization and Discrete Surface Wavelet Transform

Nineteen volumetric images were acquired using a CBCT scanner (Sirona, Bensheim, Germany) with a voxel resolution of $0.287mm \times 0.287mm \times, 0.287mm$. Thresholding segmetation was first applied to each of these volumetric images to obtain bone images. These bone images were recovered by hand-segmentation slice by slice. Then, Marching Cube Algorithm was applied on each of recovered bone images to calculate surfaces (meshes) of bone images. These bone surfaces are the ground truths for physical skeleton surfaces. We will use these 19 ground truths of bone surfaces to validate the results in Section 4 and generate landmark points in later stage.

The base of the shape is identified by first manually pinpointing 9 anatomical control landmark points [Fig. 1 (a)] on the ground truth of bone surfaces and then determining the boundaries (shortest distance) of shape Fig. 1 (b). Patch decomposition is performed by dividing the shape into 4 topological patches. Then, these patches are extracted from the bone surfaces to form a shape [Fig. 1 (c)]. Once the patches are customized, we need to find their corresponding planar domains in order to calculating the parameterization mapping. Since each of the patches is defined by control landmark points, the planar domain for each of the patches can be defined as a polygon by those control landmark points. The parameterization will be calculated using barycentric mapping [8] in the study. Mean Value Coordinate will be used to calculate the spring constants.

The Calmull-Clark subdivision [3] is performed to generate landmark points of the shape in several steps. First, the subdivision is performed on each of parameterization domains (e.g. planar K_p-polygons) of the patches to generate new

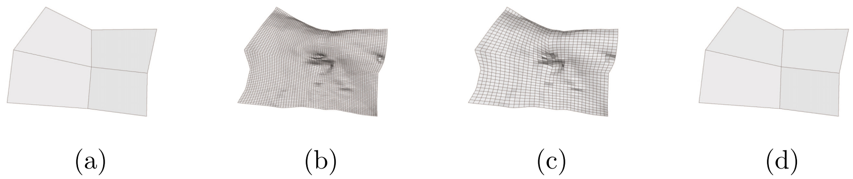


Fig. 2. (a) Base mesh for subdivision. (b) Fifth subdivision. (c) Scaling coefficients after one decomposition of DSWT. (d) Scaling coefficients after five decompositions of DSWT (coarsest mesh).

points. Then, the remeshed patches with subdivision connectivity are obtained by inverse mapping these points on the parameterization domains onto the original patches. The same number of subdivisions is used for each of the patches. Finaly, the remeshed shape with regularized landmark points can be obtained by stitching these remeshed patches. Fig. (2)(a) and (b) shows the base mesh and the remeshed stitched shape. 4225 landmark points are generated for each of the training datasets.

2.2 Wavelet Distribution Model (WDM)

Once the landmark points in each of the training shapes are generated using Calmull-Clark subdivision, these training shapes will be used to build WDM. Let S_i be the set of 4225 landmark points in ith shape in the image space. All the shapes are transformed to a common coordinate of model space using Procrustes Analysis [7] to minimize their squares of distance. Let $\tilde{S}_i$ be the transformed shape in the model space. Assume $\mathbf{x}_i$ is the $3n$-dimensional shape vectors formed by concatenation of the coordinates of all transformed landmark points in $\tilde{S}_i$.

Ths discrete surface wavelet transform (DSWT) [Fig. (2)(c) and (d)] of the remeshed shape S_i is implemented by using lifting scheme of biorthogonal wavelet transform. [1]. The construction of this DSWT is based on Calmull-Clark subdivision of arbitray two-manifold topology. Let DSWT of $\mathbf{x}_i$ be $\mathcal{W}(\mathbf{x}_i) \equiv \{\mathbf{s}_{i,0} \cup \mathbf{w}_{i,l}, l = 0, 1, \ldots, J\}$, For simplicity of notatoin, all the scaling coefficients $\mathbf{s}_{i,0}$ (base mesh) will be viewed as wavelet coefficients at (-1)-scale,i.e. $\mathbf{w}_{i,-1}$. PCA is performed on $\mathbf{w}_{i,l}$ over all the shapes to obtain the matrices $\mathbf{P}_l$ of eigenvectors. A set of wavelet coefficients $\tilde{\mathbf{w}}_l$ of a shape at specific scale l can be generated by a shape parameter $\tilde{\mathbf{b}}_l$

$$\tilde{\mathbf{w}}_l = \bar{\mathbf{w}}_l + \mathbf{P}_l \tilde{\mathbf{b}}_l \tag{1}$$

Similarly, the wavelet coefficients $\mathbf{w}_l$ of any shape at scale l can be approximated by projecting $\mathbf{w}_l$ onto the subspace of $\mathbf{P}_l$

$$\mathbf{w}_l \approx \bar{\mathbf{w}}_l + \mathbf{P}_l \mathbf{b}_l \tag{2}$$

with

$$\mathbf{b}_l = \mathbf{P}_l^T (\mathbf{w}_l - \bar{\mathbf{w}}_l) \tag{3}$$

3 A Novel Segmentation Algorithm

In the study, the training shape for statistical shape model (SSM) is exstracted from the outer surface in the skull model and is a shape-customized open surface with closed boundary. It is a partial shape in the skull model. We observe that Active Shape Model (ASM) and Wavelet Active Shape Model(WASM) may become unreliable in order to recognize the corresponding partial shape in the skull model. In the follwing, we propose a novel algorithm called base invariant wavelet active shape model (BIWASM) and an initialization method called customized wavelet base initialization (CWBI) to overcome this problem.

Let S_b be the base formed by 9 selected control landmark points by using user interaction on slices of segmented images. The selection criterion is the same as that of control landmark points demonstrated in Fig. 1 (a). Assume $\mathbf{y}_b$ is the vector formed by concatenation of the coordinates of all the points in S_b. Define

$$f(\mathcal{T}, \mathbf{b}) \equiv \|\mathbf{y}_b - \mathcal{T}\Big(\tilde{\mathbf{x}}(\mathbf{b}_l, \forall l)\Big)\|^2 \tag{4}$$

where $\tilde{\mathbf{x}}(\mathbf{b}_l, \forall l) \equiv \mathcal{C}\Big(\mathcal{W}^{-1}(\bar{\mathbf{w}}_l + \mathbf{P}_l\mathbf{b}_l, \forall l)\Big)$, and $\mathcal{C}$ is the operator extracting those corresponding control landmark points from a shape in the model space. To generate a better initial shape, we need to calcuate $\mathbf{b}_l$ and $\mathcal{T}$ by minimizing (4). $f(\mathcal{T}, \mathbf{b})$ is a non-linear function. We can use simulated annealing approach to solve it. However, these optimal search approach is computational expensive in 3D. Therefore, in this work, we propose a simple and efficient approach to estimate an initial shape. By using the model parameter $\bar{\mathbf{w}}_{-1}, \mathbf{P}_{-1}$ at base scale in (1), an $\mathbf{y}_b$ can be approximated by calculating $\mathbf{b}_{-1}$ and a transformation $\mathcal{T}_b$. Since the construction of DSWT is based on the structure of Calmull-Clark subdivision, it can be observed in Fig. 2(d) that $\mathbf{w}_{-1}$ has the same mesh structure as that in Fig. 2(a) or the same patch structure in Fig. 1(b). $\mathcal{T}_b$ is used to define a transformation between the model space and image space. A new base in the model space is defined by

$$\mathbf{w}_{-1,b} \equiv \mathcal{T}_b^{-1}(S_b) \tag{5}$$

$\mathbf{w}_{-1,b}$ cannot be exactly expressed by $\mathbf{b}_{-1}$. By combining local details using the mean wavelet coefficients $\{\bar{\mathbf{w}}_l, l \geq 0\}$ of WDM, the initial shape can be constructed by

$$\tilde{\mathbf{y}}^{(0)} = \mathcal{T}_b(\mathcal{W}^{-1}(\mathbf{w}_{-1,b} \cup \{\bar{\mathbf{w}}_l, l \geq 0\})) \tag{6}$$

This new initial shape can be interpreted as a controlled base shape attached with mean local details of WDM. $\mathcal{T}_b$ will play an essentail role in the following proposed BIWASM.

The shape in our study is different from the ones used in [6], [2], [5]. It is a partial shape in skull model and is a shape-customized open surface with closed boundary. However, ASM and WASM algorithm are not capable of constraining the evolving shape. Therefore, we design a new WASM by fixing the transformation between model space and image space. $\mathcal{T}_b$ in (5) is used to define this invariant transformation for the second step of WASM. To keep the shape contrained by the control landmark points S_b, the base shape $\mathbf{w}_{-1,b}$ in the model

Algorithm 1. Customized Wavelet Base Initialization (CWBI)

```
$\mathcal{T}_b \leftarrow (S_b, \mathbf{b}_{-1}, \bar{\mathbf{w}}_{-1})$
$\mathbf{w}_{-1,b} \leftarrow \mathcal{T}_b^{-1}(S_b)$
$\mathbf{x} \leftarrow \mathcal{W}^{-1}(\mathbf{w}_{-1,b} \cup \{\bar{\mathbf{w}}_l, l \geq 0\})$
$\tilde{\mathbf{y}}^{(0)} \leftarrow \mathcal{T}_b(\mathbf{x})$
```

Algorithm 2. Base Invariant Wavelet Active Shape Model (BIWASM)

```
$(\tilde{\mathbf{y}}^{(0)}, \mathcal{T}_b, \mathbf{w}_{-1,b}) \leftarrow (S_b, \mathbf{b}_{-1}, \bar{\mathbf{w}}_{-1})$; use Algorithm 1 to calculate the initial shape and the transformation
while Until convergence do
    $\mathbf{y}^{(k)} \leftarrow \tilde{\mathbf{y}}^{(k)}$; calculate a candidate shape by examining the neighboring region of each of the landmark points, and the corresponding control landmark points in $\mathbf{y}^{(k)}$ are replaced with $S_b$.
    $\mathbf{x}' \leftarrow \mathcal{T}_b^{-1}(\mathbf{y}^{(k)})$; calculate inverse transformatoin of $\mathbf{y}^{(\mathbf{k})}$
    $\mathbf{w}_l \leftarrow \mathcal{W}(\mathbf{x}')$; DSWT.
    $\mathbf{b}_l \leftarrow \mathbf{P}_l^T(\mathbf{w}_l - \bar{\mathbf{w}}_l), l \geq 0$; calculate the shape parameter fitting $\mathbf{w}_l$ using (3).
    Apply the constraints on $\mathbf{b}_l, l \geq 0$.
    $\mathbf{w}_l \leftarrow \bar{\mathbf{w}}_l + \mathbf{P}\mathbf{b}_l, l \geq 0$; generate wavelet coefficients in the model space using (1)
    $\mathbf{x} \leftarrow \mathcal{W}^{-1}(\mathbf{w}_{-1,b} \cup \{\mathbf{w}_l, l \geq 0\})$; inverse DSWT.
    $\tilde{\mathbf{y}}^{(k+1)} \leftarrow \mathcal{T}_b(\mathbf{x})$;
end while
```

space and S_b in the image space will be unchanged during the iteration. Image feature model is also calculated using first, second, third order derivative of image profile along each of the landmark points. The proposed algorithm is summarized in **Algorithm 2**.

4 Valiations

Nineteen sets of CBCT images were used for the validation. Their ground truths of bone surface were manually established as described above. They served as a control group. The outer surfaces of anterior wall of maxilla was segmented using our BIWASM with CBWI approach. The same images were also segmented using ASM [4] with Registration-Based Initialization (RBI), WASM with RBI, and WASM with CWBI, respectively. They all served as an experimental group.

4.1 Data Preparation

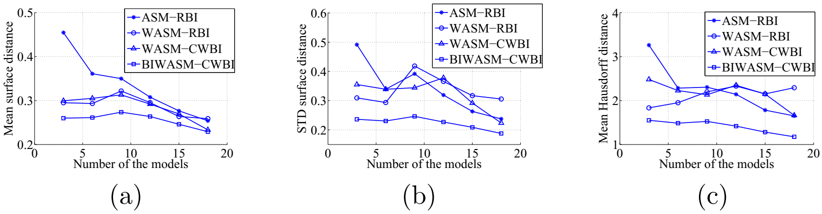
The segmentation dataset, referred as the datasets to test segmentation approaches, were labeled $D_i, i = 1, 2, \ldots, 19$. Each of the segmentation datasets D_i consisted of a set of CBCT volumetric images (target datasets) and its corresponding ground truth of bone surface. The model datasets, referred as the training datasets in the base invariant active shape model, were defined by 19 shapes and 19 sets of CBCT volumetric images. They were labeled $M_i, i = 1, 2, \ldots, 19$ in the same order.

Once N model datasets were built as training datasets, one segmentation (target) dataset, other than N model datasets, was used to compare our developed approaches to the three traditional approaches. The preparation of target dataset was completed in the following 3 steps. *Step 1* was landmark digitization. Nine control landmarks were digitized interactively for initialization. *Step 2* was initialization. The digitized control landmarks were used to create 2 initial shapes using 2 initialization methods: RBI and our newly developed CWBI. RBI was used to register these selected control landmarks of the shapes in the image space and their corresponding landmarks of mean shape in the model space, and to transform the mean shape in the model space into the image space. The resulted initial shape served as the input of ASM and WASM. CWBI was used to calculate a transformation between the base formed by those 9 control landmark points and a wavelet base in the model space. It was also used to add the mean local details using WDM to obtain an initial shape. The resulted initial shape served as the input of WASM and BIWASM. *Step 3* was to calculate the final shapes. This step produced 4 kinds of final shapes: ASM-RBI, WASM-RBI, WASM-CWBI, and BIWASM-CWBI.

Once the datasets were prepared, the final step was to compare the ground truth to the final shapes generated by different approaches. It was done by calculating surface deviations, the closest distances, and Hausdorff distance between the ground truths and the final shapes generated in the third step. Therefore, there were 4225 surface distances and one Hausdorff distance produced from each target dataset. The validation was achieved by two sets of comparisons. The first set of the comparisons was to detect the variabilities amongst 4 approaches when the number of training datasets was static, while the second set of comparisons was to detect the variabilities amongst 4 approaches when the number of training datasets was dynamic. In addition, the computational times were also compared amongst the 4 approaches.

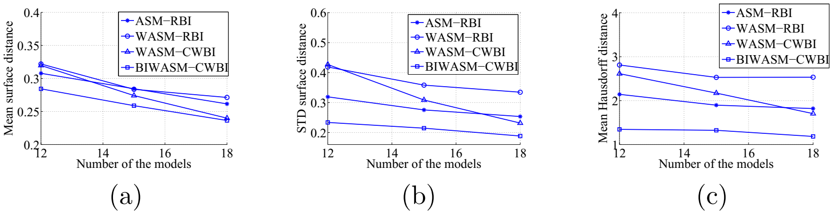
The first set of comparisons was conducted using leave-one-out arrangement (cross validation). Six groups (total of 69) leave-one-out experiments were conducted: $i = 1, 2, \ldots, 19$ (19 experiments), $i = 1, 2, \ldots, 16$ (16 experiments), $i = 1, 2, \ldots, 13$ (13 experiments), $i = 1, 2, \ldots, 10$ (10 experiments), $i = 1, 2, \ldots, 7$ (7 experiments), $i = 1, 2, \ldots, 4$ (4 experiments). The dataset was randomly selected using SPSS software. In each experiment, the target dataset was excluded from the training datasets. For example, in the second group, the experiment of the datasets $i = 1, 2, \ldots, 16$ was conducted using $M_i, i = 1, 2, \ldots, 11, 13, \ldots, 16$, and the target D_{12} dataset was excluded from the training dataset. After final shapes were generated by four approaches (ASM-RBI, WASM-RBI, WASM-CWBI, BIWASM-CWBI), they were compared to their ground truths. In each of the six groups of experiments, the mean and standard deviation of surface distances were calculated over 80275, 67600, 54925 , 42250, 29575, and 16900 surface distances, respectively. The mean Hausdorff distances was also calculated over the 19, 16, 13, 10, 7 and 4 final shapes, respectively.

The second set of comparisons was conducted by using 13 segmentation datasets $D_i, i = 1, 2, \ldots, 13$ and varying the number of model datasets by 12,



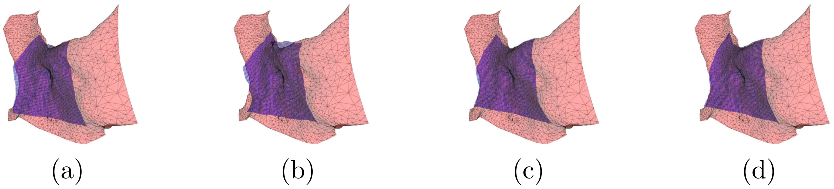
(a) (b) (c)

Fig. 3. (a), (b), and (c) are mean surface distances, standard deviation of surface distances, and mean Hausdorff distances in the first set of the comparisons



(a) (b) (c)

Fig. 4. (a), (b), and (c) are mean surface distances, standard deviation of surface distances, and mean Hausdorff distances in the second set of comparisons



(a) (b) (c) (d)

Fig. 5. The ground truth (red) and the final shapes in ASM-RBI, WASM-RBI, WASM-CWBI and BIWASM-CWBI for a single dataset

15, and 18. The dataset was also randomly selected using SPSS software. Three groups of the datasets were used to conduct 39 experiments. Again, the target dataset was excluded from the training datasets. In each group, 13 experiments were conducted using each segmentation approach, respectively. Means and standard deviations of surface distances were calculated over 54925 (13×4225) surface distances, respectively. The mean Hausdorff distances were also calculated over the 13, 13, and 13 final shapes, respectively.

4.2 Results

The results (Fig. 3, 4) showed that our BIWASM-CWBI approach outperformed the others in every single experiment. In both sets of comparisons, the largest mean surface distance was 0.26mm, standard deviation was 0.2mm, and

Table 1. The Computation Times in the First Experiment

ASM-RBI	WASM-RBI	WASM-CWBI	BIWASM-CWBI
164s	575s	618s	205s

Hausdorff distance was 1.6mm. It also indicated that the more accurate result was achieved with more training dataset. The results indicated that our BIWASM-CWBI approach was capable of capture the outer surface of thin bones (1mm) in the skull model. We noted that in the first sets of comparisons, the curves were raised when 9 sets training models were employed. This might due to poor image quality of the target dataset. This was confirmed later by our visual inspection Fig. 5.

Finally, the computational times of the 4 approaches are presented in Table 1. This was calculated in the 1st set of experiment based on D_i and $M_i, i = 1, 2, \ldots, 16, 18, 19$. The computer was Intel i7 2.8Hz with 4G RAM. The result revealed that the computational time of our approach was comparable with that of ASM-RBI and significantly shorter than WASM-RBI and WASM-CWBI.

5 Validation

References

1. Bertram, M., Duchaineau, M.A., Hamann, B., Joy, K.I.: Generalized B-spline subdivision-surface wavelets for geometry compression. IEEE Transactions on Visualization and Computer Graphics 10(3), 326–338 (2004)
2. de Bruijne, M., van Ginneken, B., Viergever, M.A., Niessen, W.J.: Adapting Active Shape Models for 3D Segmentation of Tubular Structures in Medical Images. In: Taylor, C.J., Noble, J.A. (eds.) IPMI 2003. LNCS, vol. 2732, pp. 136–147. Springer, Heidelberg (2003)
3. Catmull, E., Clark, J.: Recursively generated B-spline surfaces on arbitrary topological meshes. Computer-Aided Design 10, 350–355 (1978)
4. Cootes, T., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models-Their Training and Application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
5. Essafi, S., Langs, G., Deux, J.F., Rahmouni, A., Bassez, G., Paragios, N.: Wavelet-driven knowledge-based MRI calf muscle segmentation. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 225–228 (June 2009)
6. Essafi, S., Langs, G., Paragios, N.: Left Ventricle Segmentation Using Diffusion Wavelets and Boosting. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part II. LNCS, vol. 5762, pp. 919–926. Springer, Heidelberg (2009)
7. Goodall, C.: Procrustes Methods in the Statistical Analysis of Shape. Journal of the Royal Statistical Society 53(2), 285–339 (1991)

8. Hormann, K., Lévy, B., Sheffer, A.: Siggraph Course Notes Mesh Parameterization: Theory and Practice. In: SIGGRAPH 2007 Course Notes, pp. 1–122. ACM, New York (2007)
9. Naitoh, M., Hirukawa, A., Katsumata, A., Ariji, E.: Evaluation of voxel values in mandibular cancellous bone: relationship between cone-beam computed tomography and multislice helical computed tomography. Clin. Oral Implants Res. (2009)
10. Scarfe, W.C., Farman, A.G., Sukovic, P.: Clinical applications of cone-beam computed tomography in dental practice. J. Can. Dent. Assoc. 72(1), 75–80 (2006)
11. White, S.C., Pharoah, M.J.: Oral Radiology: Principles and Interpretation, 4th edn. Mosby, St. Louis (2000)

Searching Structural Blocks by SS Exhaustive Matching

Virginio Cantoni[1], Alessio Ferone[2], Ozlem Ozbudak[3] and Alfredo Petrosino[2]

[1] University of Pavia, Department of Electrical and Computer Engineering,
Via A. Ferrata, 1, 27100, Pavia, Italy
virginio.cantoni@unipv.it
[2] University of Naples Parthenope, Department of Applied Science,
Centro Direzionale Isola C4, 80133, Napoli, Italy
{alfredo.petrosino,alessio.ferone}@uniparthenope.it
[3] Istanbul Technical University
Department of Electronics and Communication Engineering, Ayazaga Campus,
34469, Maslak, Istanbul, Turkey
ozbudak@itu.edu.tr

Abstract. This paper presents motif retrieval from a macromolecule or a protein by using structure comparison in 3D through an exhaustive matching analysis of secondary structures. The comparison is based on three parameters: midpoint distance (Md), axis distance (Ad) and angle (φ) related to a couple of SSs in 3D space. The barycenter of the motif is assigned as Reference Point (RP) and in order to find the RP related to every possible motif (instance) in the macromolecule a voting process is performed. The searched motif is compared with all possible instances having the same number of motif SSs in the macromolecule and gives a vote to the candidate barycenter for every correspondence. The point, which has the maximum number of votes, is determined as candidate RP. In this paper motifs composed by four and five secondary structures are searched. Experimental results show a good accuracy in determining the RP and hence in the retrieval of the searched motif.

Keywords: structural motif search, protein motif retrieval, protein structure comparison, exhaustive matching, protein secondary structure.

1 Introduction

Proteins are crucial molecules in biological phenomena because they form much of the functional and structural machinery in every organisms. The function performed by each protein is determined by its spatial structures that can be described at various level of detail, ranging from atomic coordinates, through vector approximations, to SS elements. Protein structure comparison is an important issue that helps biologists understand various aspects of protein function and evolution. Indeed the 3D fold has a major effect on the ability of a protein to bind other proteins or ligands. Therefore, similarity analysis in terms of protein structure is very important in order to uncover the role of an unknown protein.

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 57–69, 2013.

Comparison of protein structures is also essential for estimating the evolutionary distances between proteins and protein families.

The structural comparison problem in a protein structure retrieval system has been studied in several computational biology literatures. Can et al. [3] present a new method for conducting protein structure similarity searches and applies differential geometry knowledge on protein 3D structure for extracting signatures such as curvature, torsion and SS type. Camoglu et al. [2] build an indexing structure based on SS elements triplets by using R-tree in order to find similarities in a protein structure database. Chionh et al. [9] propose the SCALE algorithm to compare protein 3D structures based on angle-distance matrices that utilizes angles and distances between SS elements. Chi et al. [8] design a fast system for protein structure retrieval by using image based distance matrices and a multidimensional index. Zotenko et al. [17] propose an approach to speed up protein structure comparison by mapping a protein structure to a high-dimensional vector and approximating structural similarity by distance between the corresponding vectors. Krissinel et al. [14] describe the Secondary Structure Matching (SSM) algorithm of protein structure comparison in 3D, which includes an original procedure of matching graphs built on the protein's SS elements, followed by an iterative 3D alignment of protein backbone C_α atoms. Cantoni et al. [4, 5] made a study for retrieving structural motifs by using Hough transform and range tree. They also retrieved the Greek Key motif, which is formed by four SSs, from the protein files by using the GHT [6, 7].

A structural motif is a 3D structural element which appears in a variety of molecules and usually consists of just a few secondary structures. Several motifs packed together to form compact, local, semi-independent units are called domains. The size of individual structural domains varies from between about 25 up to 500 amino acids, but the majority, 90%, has less than 200 residues with an average of approximately 100 residues. The protein family is a group of evolutionarily related proteins, that have a common ancestor and typically have similar 3D structures, functions, and significant sequence. Note that it is also often used the term super-*, where * can stand for motif, or domain, or family, or fold, or class [5]. Many methods have been proposed for defining protein SSs, but the Dictionary of Protein Secondary Structure (DSSP) [13] method is the most commonly used. This method classifies eight types of SSs, even though in most cases only the three dominant configurations are considered: helices include 3_{10}-helix, α-helix and π-helix; sheets or strands include extended strand (in parallel and/or anti-parallel β-sheet conformation); finally, coils include hydrogen bonded turn, bend, and amino acid residues which are not in any of the previous types. The structural analysis for protein recognition and comparison is conducted mainly on the basis of the two most frequent components [11]: the helices and the strands.

This paper is organized as following. In Section 2, we introduce briefly the GHT and then explain the exhaustive matching algorithm that we adopt for motif retrieval. In Section 3, we represent the experiments and their results. Finally, we conclude this paper with possible future works in Section 4.

2 Methods

2.1 Generalized Hough Transform

In this paper an algorithm based on GHT is used. This transform is an extracting method using coordinate transformation [12]. It was introduced by P.V.C. Hough in 1962 and patented by IBM. Hough used angle-radius parameters exclusively for retrieving the straight lines. HT was extended for extracting circles by R.O. Duda and P.E. Hart in 1972 [10] and for retrieving parabolas by H. Wechsler and J. Sklansky in 1973 [16]. Later it was generalized as GHT by Ballard for retrieving arbitrary shapes [1]. Basically, the original HT is a voting process where each contour point detected in the image votes for all possible patterns passing through that point. As an example in the implementation to detect straight lines, votes are accumulated in an array $A(\rho, \theta)$, where θ is the angle made by the normal to the straight line with the x-axis and ρ is the perpendicular distance of the straight line from the origin. The representation of the straight line in (ρ, θ) form is

$$x \cos\theta + y \sin\theta = \rho \tag{1}$$

This accumulator array $A(\rho, \theta)$ is called the Hough Space (HS). The number of votes for each cell in $A(\rho, \theta)$ represents the number of pixels in the searched pattern extracted from the image. In this process each pixel in the image space is mapped to a sinusoidal curve of Eq. 1 in the HS. So HT is a transformation from a point to a curve [15]. In GHT arbitrary shapes are represented in the HS which consists of the parameters of the rigid motion - in 2D (x, y, θ, s) x, y representing translation, θ rotation and s a scaling factor. For each 'evidence' extracted from the image, like in the original Hough approach, a mapping rule is defined which determines the value of the parameters of rigid motion (locus of points in HS) compatible with the 'evidence'.

In this paper the GHT is exploited for comparison and search of structural similarity between a given motif or domain or entire protein and the proteins of a database, like, for instance, the PDB [18]. Note that, if the searched structure is just a component of a protein (like a structural motif or a domain) the same algorithm supports the detection and the statistical distribution of these components.

2.2 Exhaustive Matching

This algorithm directly matches the motif and all possible instances in the macromolecule and uses the couple parameters in order to match. The number of couples in motif is given by Eq. 2,

$$C = (m, 2) = \frac{m!}{(m-2)!2!} \tag{2}$$

where m is the number of SSs into the motif. For every motif couple Md, Ad and φ are calculated. In Fig. 1 a couple of SSs (A and B) is represented. The local reference system is highlighted together with the three quoted parameters and the corresponding RP position. Considering midpoint coordinates, the barycenter of the motif is determined and selected as RP. In the voting process we will vote on the barycenters of the instances in the macromolecule. In this algorithm let N, m, M be the number of SSs in the macromolecule, the number of SSs in the motif and the number of instances in the macromolecule respectively, i.e. M is the m-combinations in N. So, it can be computed as:

$$M = (N, m) = \frac{N!}{(N-m)!m!} \tag{3}$$

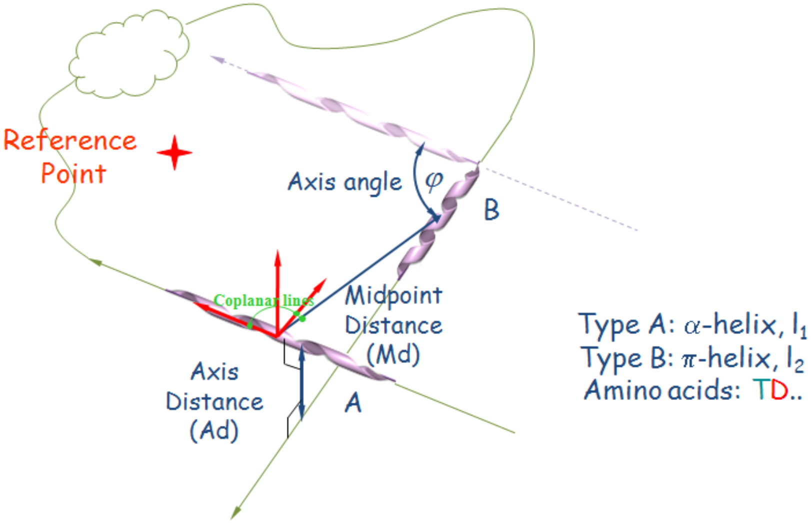


Fig. 1. Representation of the local reference system and the couple parameters: Md, Ad, φ

For every instance, C Md, C Ad, C φ (totally Cx3 values) and barycenters location are determined. Then every parameter is compared to the expected values considering a heuristic error rate, 1%. If the compared parameters are compatible, a vote is given to the barycenter related to that instance. Tab. 1 describes a sketch of this algorithm, as it can be easily seen, the computational complexity is of the order $O(CM) \cong O(N^m)$. In Fig. 2 a graphic sketch of this process is given. Here, the aim is at searching the motif model on the top right in the macromolecule on the left. To do this, couple parameters are used. In this figure only the parameter Md is represented. Md values ($d1$, $d2$, $d3$) for the three couples in the motif are stored in the RT. If the motif couple parameters are equal to macromolecule couple parameters the RP is determined according to the mapping rule related to that motif couple. On the top of Fig. 2 a complete instance of the model is present, so the correspondent RP position will gather three contributions (only one is graphically shown). Instead on bottom right an instance constituted of two SSs is present: only one contribution is given in the correspondent RP' position.

Later a 3D vote space is built and then this space is scanned by using a KxKxK cubic mask for neighbors vote cumulation. In this experimentation K has been limited to 3. After the scanning process the point/points having the highest votes is/are determined and is/are considered as candidate RP of the motif.

Table 1. Sketch of the algorithm for the structural block exhaustive matching search

```
Input  : Protein .nss file; $N$: number of protein SSs; $m$: number of motif SSs
Output : Locations of candidate motifs in the accumulator $A_{RP}$, representing
         the parameter space.
Select randomly $m$ SSs among the $N$ SSs of the protein to create the motif
         model.
Find the motif barycenter location as $RP$.
Calculate the number of couples in the motif: $C = (m, 2)$.
for $i = 1, C$
    Compute the three parameters: $Md_i$, $Ad_i$ and $\varphi_i$
Compute $M = (N, m)$.
for $j = 1, M$
    Compute $Md_j$, $Ad_j$ and $\varphi_j$ parameters.
    Compute $RP_j$ location.
    for $i = 1, C$
        if ($Md_j$, $Ad_j$, $\varphi_j$ equal to $Md_i$, $Ad_i$, $\varphi_i$) then $A_{RPj} = A_{RPj} + 1$
```

3 Experiments and Results

The proposed algorithm consists of an exhaustive matching for searching a general motif in the macromolecule. Firstly a motif model is created by using SSs of a given macromolecule, e.g. for creating five-SS motif, in our experimentation, five SSs of the macromolecule are selected randomly and used. Later two peculiar motifs (the Greek Key motif) and four random motifs are selected as four-SS motif. For the searching algorithms, Md, Ad and φ parameters are used as comparison parameters. Md is the Euclidean distance between middle points of two SSs, Ad is the shortest distance between two SSs axis and φ is the angle between two SSs translated to present common extreme (see Fig. 1).

In a first set of the experiments 1FNB, 4GCR and 7FAB proteins were used as macromolecule for searching a motif composed by five SSs. These proteins are shown in the lower side of Figs. 3, 4 and 6. From these proteins five SSs

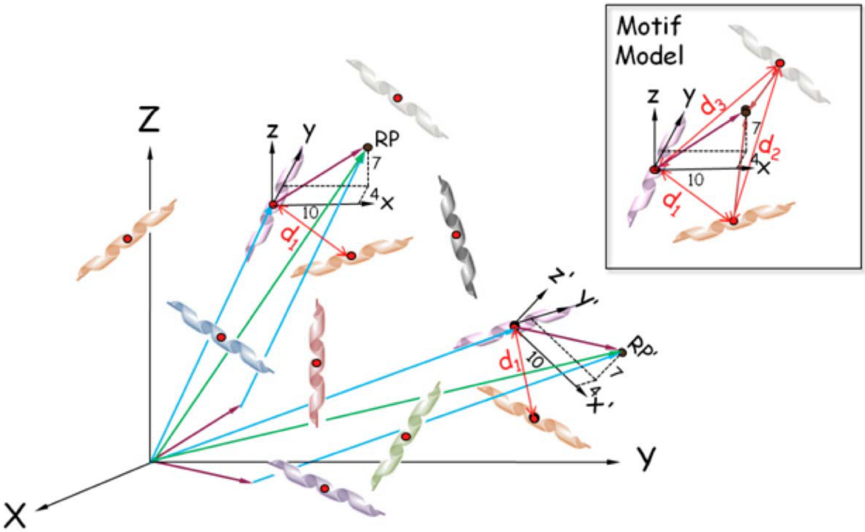


Fig. 2. The voting process for a few couple of helices. On the top right a sketch of a motif having three SSs and on the left RP location of a compatible couple.

were selected randomly to create the motif model and totally six different motifs (two motifs for each protein) were used for testing. Figures 3 and 4 represent two motifs from 1FNB and 7FAB proteins and their parameters space. Table 2 shows the number of SSs in the proteins, the number of five-SS instances in these proteins and the number of couples in the motif.

Table 2. Proteins and a few important parameters

PDB ID	#SSs in the protein	#SSs in the motif	#Instances in the protein	#Couples in the motif
1FNB	22	5	26334	10
4GCR	18	5	8568	10
7FAB	46	5	1370754	10

The motif has 10 couples so for every couple parameters 10 values were calculated, and the expected number of vote having detected all possible matches is 30. Then the algorithm in Tab. 1 was performed. A 3x3x3 cubic mask was applied to the parameter space. Then clouds were found where the votes are cumulated and the center of this cloud was determined as candidate RP. The results are represented in Tab. 3.

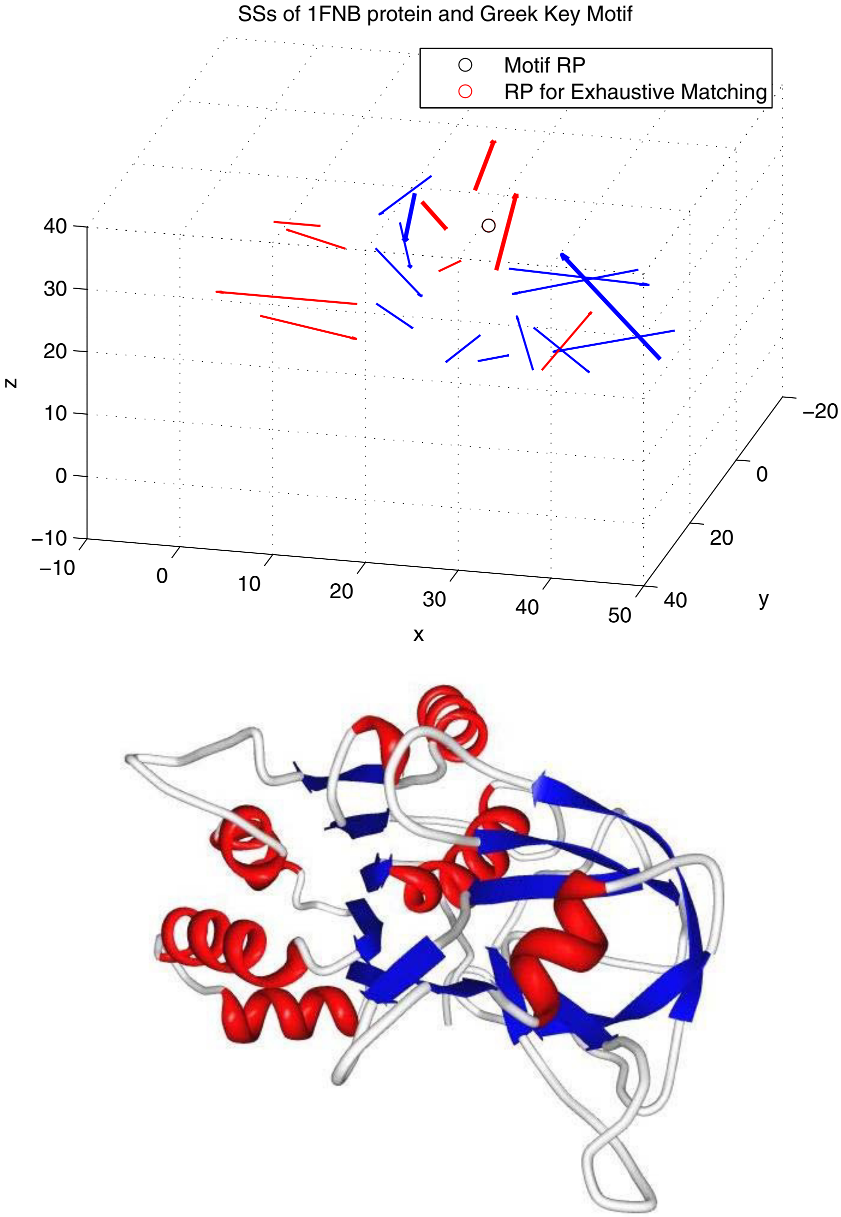


Fig. 3. SSs of the 1FNB protein. Red lines are α-helices and blue lines are β-strands. Bold lines form the five-SS motif. RP and Max. vote coordinates are almost coincident for both algorithms.

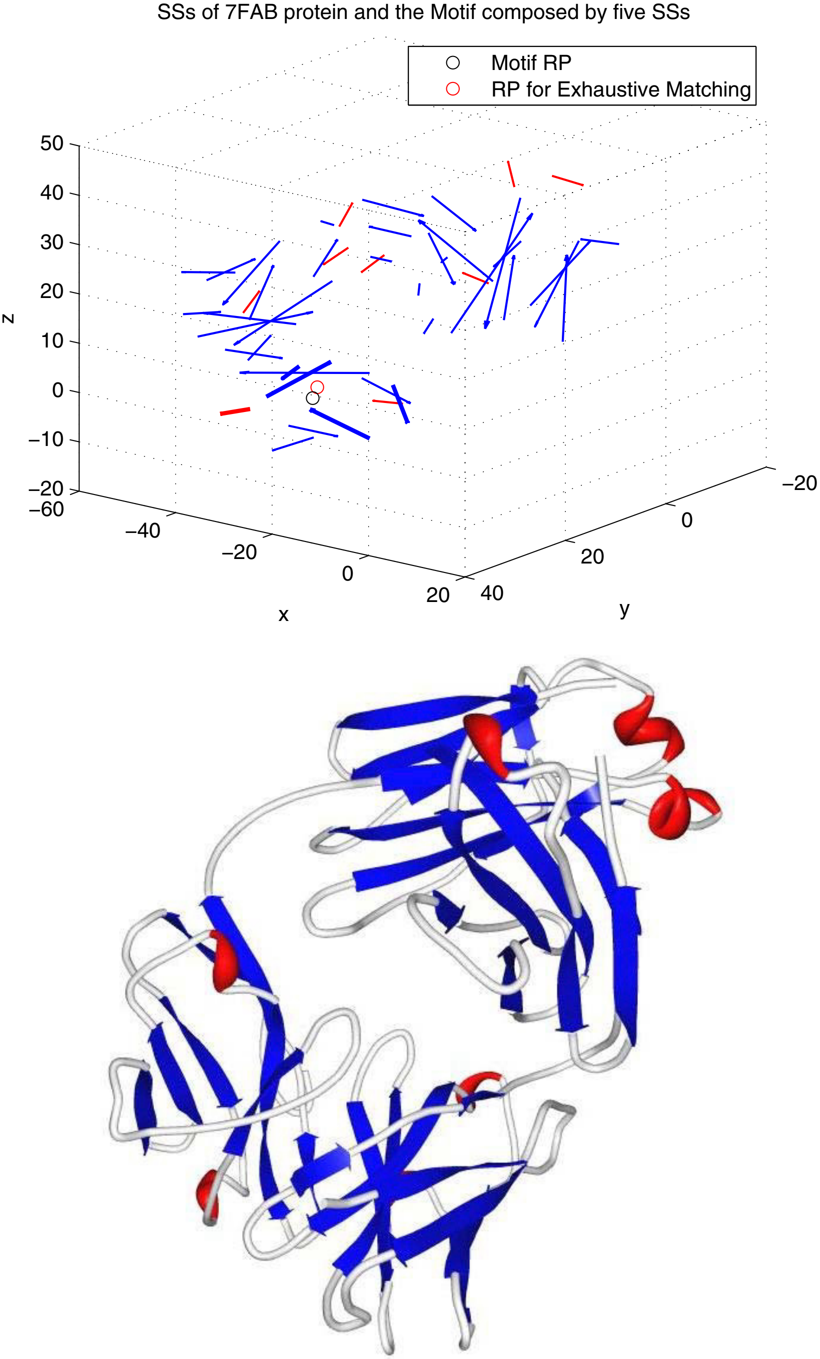


Fig. 4. SSs of the 7FAB protein. Red lines are α-helices and blue lines are β-strands. Bold lines form the five-SS motif. RP and Max. vote coordinates are almost coincident for both algorithms.

Table 3. Results for searching the motif formed by five SSs by exhaustive matching

PDB ID	Mask dim.	Motif RP	Candidate RP	#Max votes	Error rate
1FNB	3x3x3	[32.13 0.24 11.75]	[32.11 0.26 11.77]	30	0.10%
1FNB	3x3x3	[22.91 -1.61 23.34]	[22.91 -1.64 23.37]	30	0.13%
4GCR	3x3x3	[10.50 15.70 29.17]	[10.14 15.81 29.40]	36	1.27%
4GCR	3x3x3	[8.62 16.16 25.08]	[6.67 16.76 26.72]	34	8.43%
7FAB	3x3x3	[-31.88 20.51 -2.24]	[-31.37 20.09 -0.08]	30	5.95%
7FAB	3x3x3	[-24.74 17.59 24.23]	[-25.35 17.46 23.16]	33	3.19%

As shown in Tab. 3 the RP of the motif is almost coincident to the RP of the candidate instance. Here the motif is retrieved from the macromolecules 1FNB, 4GCR and 7FAB with 0.12%, 4.85% and 4.57% average error rates respectively. Note that in three cases the peaks are increased by spurious votes given by surrounding cooccurrences compatible to the admitted tolerance. Nevertheless this increasing is negligible for the block detection!

In the second set of the experiments we used the same proteins but different motifs. Here, two particular (the Greek Key motif) and four random motifs were used. Figures 5 and 6 represent the Greek Key motifs from 1FNB and 4GCR proteins and their parameters space. Table 4 shows the number of SSs in the proteins, the number of instances in these proteins and the number of couples in the motif.

Table 4. Proteins and a few important parameters

PDB ID	#SSs in the protein	#SSs in the motif	#Instances in the protein	#Couples in the motif
1FNB	22	4	7315	6
4GCR	18	4	3060	6
7FAB	46	4	163185	6

In this part of the experiments the motif has 6 couples. So expected number of vote is 18. The results with the exhaustive matching approach are shown in the Tab. 5. The RP of the motif is almost coincident to the RP of candidate instance. The Greek Key motifs and the other motifs are retrieved from the macromolecules 1FNB, 4GCR and 7FAB with 0.21%, 0.24% and 0.12% average error rates respectively. Note that also in this case there are two peaks slightly different in amplitude from the expected value. One is increased by three extra contribution, also in this case given by surrounding cooccurrences. A second is decreased by just one vote, absolutely negligible for the detection decision, due probably to a rounding calculus error.

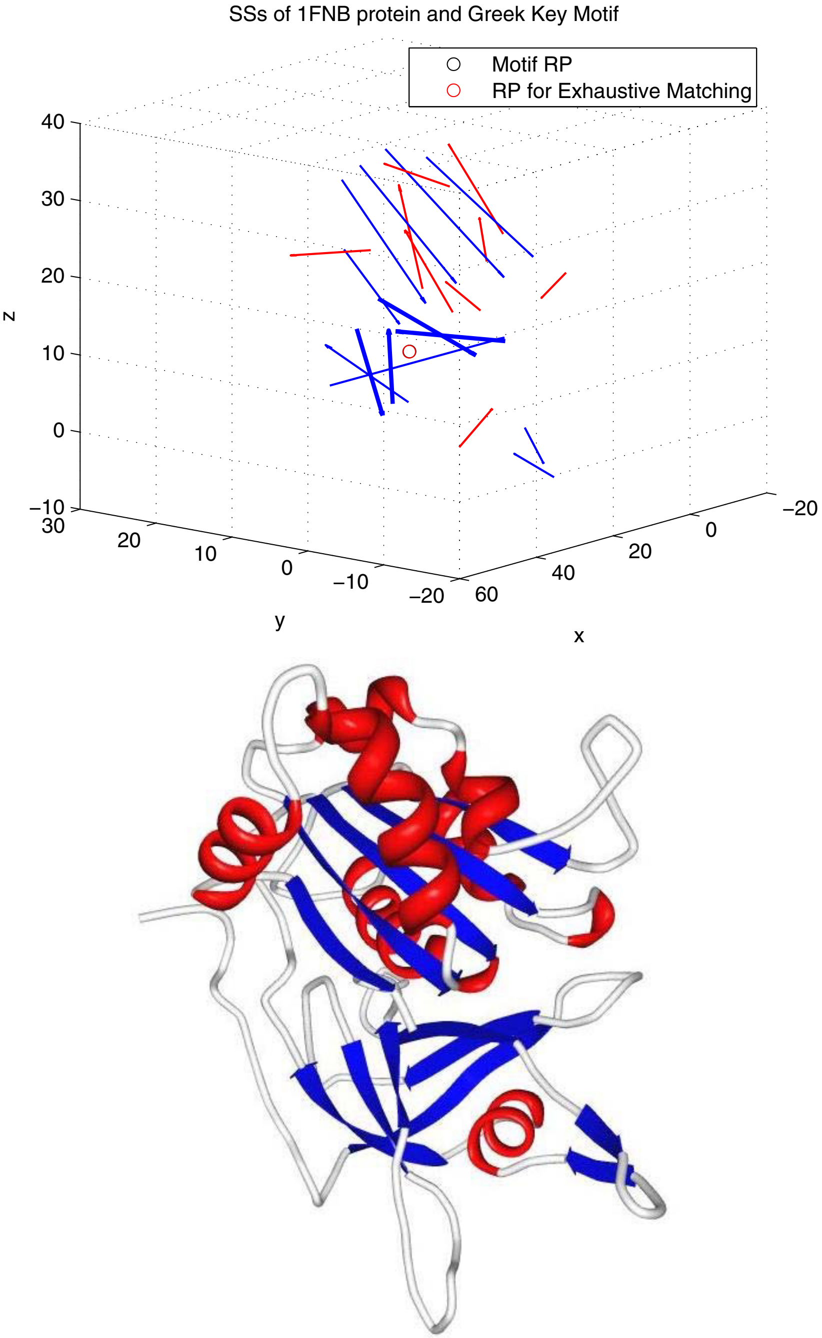


Fig. 5. SSs of the 1FNB protein. Red lines are α-helices and blue lines are β-strands. Bold lines form the Greek Key motif. RP and Max. vote coordinates are almost coincident for both algorithms.

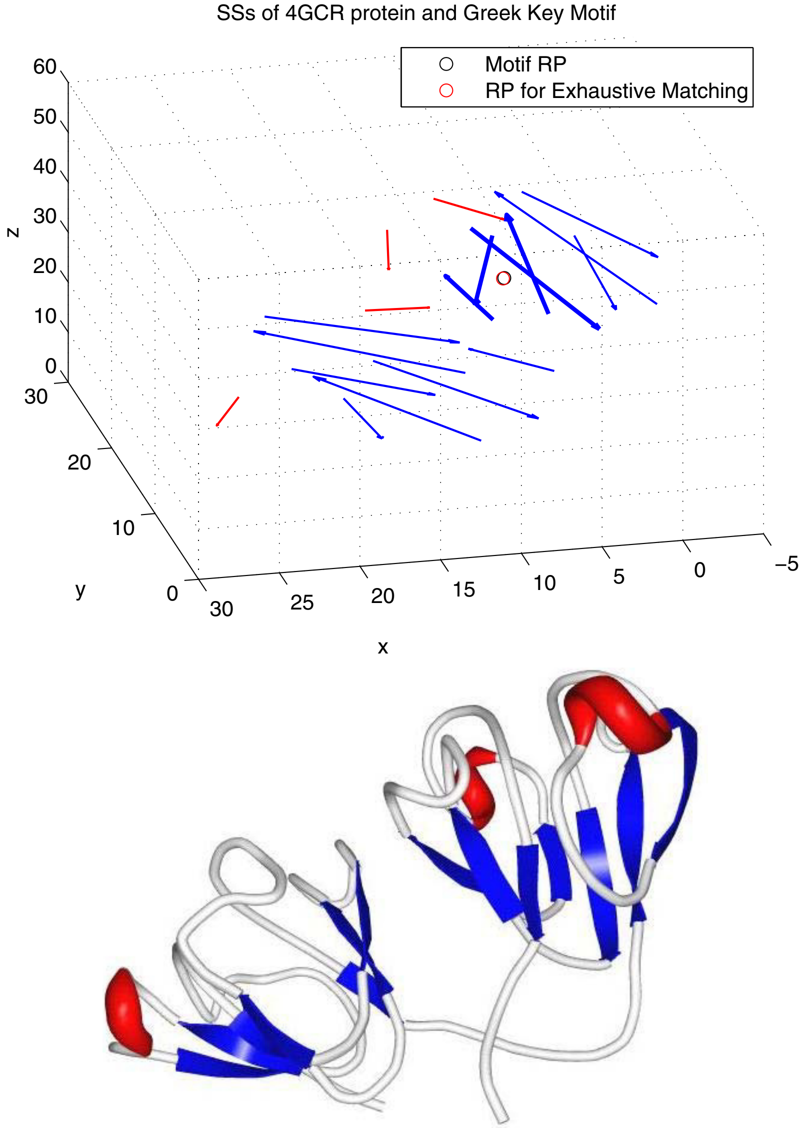


Fig. 6. SSs of the 4GCR protein. Red lines are α-helices and blue lines are β-strands. Bold lines form the Greek Key motif. RP and Max. vote coordinates are almost coincident for both algorithms.

Table 5. Results for searching the motif formed by four SSs by exhaustive matching

PDB ID	Mask dim.	Motif RP	Candidate RP	#Max votes	Error rate
1FNB	3x3x3	[31.38 1.08 11.69]	[31.40 1.12 11.66]	18	0.16%
1FNB	3x3x3	[13.81 3.40 17.11]	[13.80 3.42 17.06]	18	0.25%
4GCR	3x3x3	[7.06 14.85 34.81]	[7.14 14.84 34.74]	21	0.28%
4GCR	3x3x3	[12.85 17.20 13.52]	[12.89 17.19 13.49]	18	0.20%
7FAB	3x3x3	[-17.51 11.68 22.94]	[-17.54 11.71 22.95]	17	0.14%
7FAB	3x3x3	[-29.04 21.41 25.94]	[-29.04 21.41 25.90]	18	0.09%

4 Conclusion and Future Work

The function of a protein is determined by its spatial structure so it is important to learn structure-function relationship in the protein universe by comparing their structures and retrieving similar motifs, domains, proteins. This paper aims at retrieving a motif (formed by SSs). In order to retrieve the motif the algorithm uses exhaustive matching. Using this algorithm two sets of experiments were performed. The first set experiments consist of retrieving five-SS motif, the result shows that the RP of the instance is almost coincident to the expected RP. We can say that this algorithm is successful for retrieving a motif from the macromolecule. A second set of experiments consist of retrieving also the Greek Key motif from the 1FNB and 4GCR proteins. Also in this case the results showed that the RP of the instance having the maximum vote is almost coincident to the expected RP. For four-SSs motif retrieval the test results are quite encouraging, obviously more experiments are required for a valid statistics.

In future works, co-occurrence with primitives of more than two SSs can be pursued and also it can be experimented for different types of motifs and higher number of SSs up to structural domains. Moreover, the approach can be tested also for domain and protein comparison and search.

References

[1] Ballard, D.H.: Generalizing the Hough Transform to Detect Arbitrary Shapes. Pattern Recognition 13(2), 111–122 (1981)

[2] Camoglu, O., Kahveci, T., Singh, A.: PSI: Indexing Protein Structures for Fast Similarity Search. Bioinformatics 19(suppl. 1), 81–83 (2003)

[3] Can, T., Wang, Y.F.: CTSS: A Robust and Efficient Method for Protein Structure Alignment Based on Local Geometrical and Biological Features. In: Proc. of the IEEE Computer Society Conference on Bioinformatics, pp. 169–179 (2003)

[4] Cantoni, V., Mattia, E.: Protein structure analysis through Hough transform and range tree. Nuovo Cimento della Società Italiana di Fisica. C, Geophysics and Space Physics 35, 39–45 (2012)

[5] Cantoni, V., Ferone, A., Petrosino, A.: Protein motif retrieval through secondary structure spatial co-occurrences. Nuovo Cimento della Società Italiana di Fisica. C, Geophysics and Space Physics 35 C, 47–54 (2012)

[6] Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Structural Analysis of Protein Secondary Structure by GHT. In: 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, pp. 1767–1770. IEEE Computer Society (2012) ISBN: 9784990644116

[7] Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Search of Protein Structural Blocks through Secondary Structure Triplets. In: 3rd International Conference on Image Processing Theory, Tools and Applications, IPTA 2012, Istanbul, Turkey, October 15-18, pp. 222–226. IEEE Press (2012) ISBN: 9781467325844

[8] Chi, P.H., Scott, G., Shyu, C.R.: A Fast Protein Structure Retrieval System Using Image Based Distance Matrices and Multidimensional Index. International Journal of Software Engineering and Knowledge Engineering, Special Issue on Software and Knowledge Engineering Support in Bioinformatics, 522–532 (2004)

[9] Chionh, C.H., Huang, Z., Tan, K.L., Yao, Z.: Augmenting SSEs with Structural Properties for Rapid Protein Structure Comparison. In: Proc. of the Third IEEE Symposium on Bioinformatics and Bioengineering, pp. 341–348 (2003)

[10] Duda, R.O., Hart, P.E.: Use of the Hough Transformation to Detect Lines and Curves in Pictures. Comm. ACM 15(1), 11–15 (1972)

[11] Eisenberg, D.: The Discovery of the alpha-helix and beta-sheet, the Principal Structural Features of Proteins. Proc. of the National Academy of Sciences of the United States of America 100(20), 11207–11210 (2003)

[12] Hough, P.V.C.: Methods and Means for Recognizing Complex Patterns. US Patent 3069654 (1962)

[13] Kabsch, W., Sander, C.: Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. Biopolymers 22(12), 2577–2637 (1983)

[14] Krissinel, E., Henrick, K.: Secondary-structure Matching (SSM), a New Tool for Fast Protein Structure Alignment in Three Dimensions. International Union of Cyrstallography, Acta. Cryst. D60, 2256–2268 (2004)

[15] Naik, S.K., Murthy, C.A.: Hough Transform for Region Extraction in Color Images. In: Proceedings of ICVGIP, pp. 252–257 (2004)

[16] Wechsler, H., Sklansky, J.: Automatic Detection of Ribs in Chest Radiographs. Pattern Recognition 9, 21–30 (1977)

[17] Zotenko, E., Dogan, R.I., Wilbur, W.J., O'Leary, D.P., Przytycka, T.M.: Structural Footprinting in Protein Structure Comparison: The Impact of Structural Fragments. BMC Structural Biology 7(1), 53 (2007)

[18] http://www.rcsb.org/pdb/

Characterizing Intermediate Conformations in Protein Conformational Space

Rosanne Vetro, Nurit Haspel, and Dan Simovici

Department of Computer Science, University of Massachusetts Boston
100 Morrissey Blvd. Boston MA 02125 USA
{rvetro,nurith,dsim}@umb.edu

Abstract. In this paper we present a novel parallel coordinate based clustering method using Gaussian mixture distribution models to characterize the conformational space of proteins. We detect highly populated regions which may correspond to intermediate states that are difficult to detect experimentally. The data is represented as feature vectors of N dimensions, which are lower-dimension projections of the protein conformations. Parallel coordinates are a visualization technique that lays out coordinate axes in parallel rather than orthogonal to each other, thereby allowing patterns between pairs of axis as well as outliers to be visually identified in multi-dimensional data. We believe that the size of the resulting clusters may provide information about the likelihood of the corresponding conformations to exist as important intermediates. We tested our method on the conformational space for the enzyme Adenylate Kinase (AdK) which undergoes large scale conformational changes and used our method to detect clusters which may correspond to experimentally known intermediates. Finally, we compare our clusters with the ones generated by the K-Means clustering algorithm and discuss the advantages of our method for the problem of characterizing proteins conformational space.

Keywords: Clustering, Parallel Coordinates, Protein Conformational Search, AdK, Structural Bioinformatics.

1 Introduction

Proteins are flexible molecules that undergo structural (conformational) changes as part of their interactions with other proteins or drug molecules [1]. Changes in torsional angles may induce localized changes or large scale domain motions. Characterizing the conformational space of proteins is crucial for understanding the way they perform their function. There is promise that understanding the connection between protein structure, dynamics and function can contribute a lot to our understanding of how molecular machines function. Therefore, the question of how the structure and dynamics of proteins relate to their function has challenged scientists for several decades but still remains open.

Existing physics-based computational methods that sample the conformational space of proteins include Molecular Dynamics (MD) [2], Monte Carlo (MC) [3] and their variants, as well as approximate methods based on geometric sampling [4–7], Elastic Network Modeling [8], normal mode analysis [9], morphing [10] and more. One of the main challenges in modeling conformational changes in proteins is the difficulty in

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 70–80, 2013.

detecting intermediate structures that may correspond to transition states. These intermediate states are transient and therefore hard to detect experimentally, but they may be crucial to understanding folding, docking, binding and conformational change processes, as well as for drug design, since many times a drug is targeted as a transition state analog or to block the target molecule from undergoing a structural change. In addition, full-scale conformational search of even a medium sized protein is very demanding computationally and the conformational landscape of proteins is many times rugged and hard to navigate. Therefore, the challenging problem of fully characterizing conformational pathways in proteins still remains open. Recently, we developed a semi-coarse grained conformational search method that conducts a fast, approximate search on the conformational space of proteins undergoing large-scale domain motions [4]. While the method produced feasible conformational pathways, these pathways needed to be clustered and filtered to extract meaningful intermediate conformations.

In this work we use a variant of the above mentioned conformational search algorithm [4] to provide an approximate description of the protein conformational landscape. The algorithm runs a large number of monte-carlo like searches in the conformational space of proteins undergoing large-scale changes. By repeating the procedure a large number of times we produce a set of feasible pathways which provide a good coverage of the space.

In order to find highly populated regions which may correspond to intermediate structures, we introduce a clustering method that takes as input conformational pathways represented by a lower-dimensional projection of the protein conformational space and outputs clusters of data that can give us information about the likelihood of the existence of given structures. The method performs a statistical analysis of multi-dimensional data representing conformations. Each dimension is partitioned in a possibly different number of blocks using model-based clustering with Gaussian mixture models and the data flow between pairs of dimensions is analyzed in order to create disjoint multi-dimensional clusters of conformations and identify structures that are unlikely to be meaningful local minima as outliers.

Experimental results regarding the conformational space for the enzyme Adenylate Kinase (AdK) suggest that the combination of our conformational search and clustering method can help us detect highly populated areas in the conformational space, represented by large clusters, which may indicate the location of important intermediate structures in the protein conformational space, as demonstrated by similarity to known AdK intermediate homologs. In order to evaluate our clustering method, we compare our results with the ones generated by multiple runs of the K-means algorithm [11, 12] and present the advantages of our approach.

The paper is organized as follows. Section 2 presents the methods for protein conformational search and clustering. The methodology is evaluated experimentally in section 3. Finally, the paper is concluded in section 4.

2 Methods

2.1 Protein Conformational Search

We use a semi–coarse grained protein structure representation. The proteins are stripped of their side-chain and hydrogen atoms and represented at the backbone and C-β level

(Glycine is represented by its backbone only). We apply a semi-coarse-grained potential function to approximate the protein energy [13] and an efficient distance measure to estimate the distance between two protein structures based on the positions and angles of their secondary structure elements [4]. This measure represents each protein conformation as a feature vector whose size is the order of magnitude of the number of rigid elements in the protein, thus projecting the structures onto a much lower dimensional space than its full representation and is more "natural" to protein structures than other projections such as PCA, since PCA is limited by its linear nature.

The search algorithm used here is a variant of a Monte Carlo search that leads from one conformational state of the protein (start) to another (goal), applying successive geometric transformations to a randomly selected backbone degree of freedom of the structure while retaining only intermediate structures with an energy below a threshold [4]. We used that method to validate the results at the previous paper. In this paper we used the Monte-Carlo based search rather than a Robotics based search used in the previous work [14], since that method tends to bias the results towards the goal structure and in the present work we wanted to generate as random a sampling of the low-energy conformational space as possible. The reader is encouraged to refer to [4] for more details about the conformational search method.

2.2 Clustering Method

Today, the majority of clustering methods for multi-dimensional data incorporates metric functions that evaluate the distance between feature vectors extracted from a data-set. In this scenario, multiple dimensions are combined and are simultaneously considered according to a metric function in order to create a set of clusters.

In this paper, we propose an alternative clustering method based on parallel coordinates [15] and Gaussian mixture models [16], and argue that it is suited for providing information about the likelihood of the existence of given intermediate conformations in a protein conformational space.

To formally describe our clustering method, we introduce the following notation. The symbol $\mathscr{C}$ stands for a set of conformations represented by feature vectors in the data set, n for the conformational space dimensionality and Σ is the $n \times n$ covariance matrix of the data set where each element along the diagonal, $\Sigma[i,i]$, corresponds to the variance of dimension d_i with $1 \leq i \leq n$. The statistical information provided by Σ is used to create $\mathscr{L}$, an ordered list of dimensions. The threshold used by the algorithm to find outliers or conformations that are unlikely to exist is given by τ where $0 \leq \tau \leq 1$. It corresponds to the minimum fraction of diverging vectors that can form a new cluster, considering the total number of vectors in the original cluster from which the split occurred. Finally, B stands for the matrix containing information about the blocks of each dimension's partition. For instance, $B[i,j]$ corresponds to the block from dimension j in which the corresponding data value from conformation i is located.

Parallel coordinates is a common way of visualizing high-dimensional geometry and analyzing multivariate data. Dimensions or axis are laid out in parallel rather than orthogonal to each other. Each data value of an n-dimensional vector is positioned on the line corresponding to its axis, between the minimum (at the bottom) and the maximum (at the top) values of the axis. Points belonging to the same vector are connected

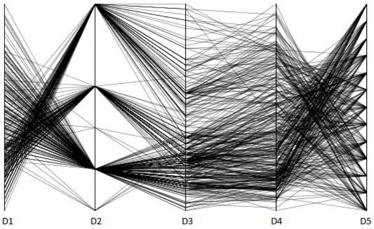


Fig. 1. Example of parallel coordinates for a 5-dimensional data set

by lines, which allows patterns between dimensions and outliers to be visually identified. For example, Figure 1 shows a 5-dimensional data set displayed as a sequence of parallel coordinates. Notice the inverse relationship between $D1$ and $D2$ and the correlation between $D2$ and $D3$: lower values of $D1$ usually imply higher values of $D2$ and vice-versa; higher values for $D2$ usually imply higher values for $D3$ and vice-versa. Likewise, we can visualize the correlation between $D3$ and $D4$ and the cross among the lines between $D4$ and $D5$. Exceptions or outliers corresponding to diverging lines that disrespect the usual behavior between dimensions can also be spotted using this technique.

The real strength of parallel coordinates is in modeling relations between variables, as discussed in [17]. Our method analyzes the variance of each dimension to model those relations. The model is simply represented using $\mathscr{L}$. The purpose of $\mathscr{L}$ is to determine the order in which the algorithm will analyze the data flow between consecutive pairs of dimensions in order to form clusters.

Given a set $\mathscr{C}$ of conformations represented by feature vectors in n dimensions, the covariance matrix Σ of the data set is generated and all dimensions are placed in $\mathscr{L}$ in increasing order of variance. We do not claim that this arrangement is optimal. The optimal ordering of the dimensions is a topic for further study.

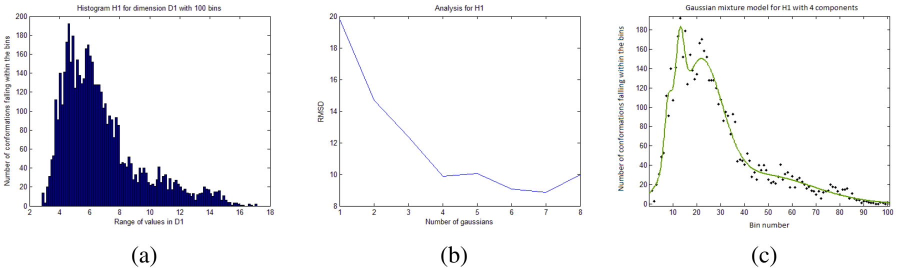
In order to assign a unique cluster to each conformation and identify outliers our method first performs a model-based clustering on each dimension separately using Gaussian mixture distribution models to estimate density. A Gaussian or normal mixture model is a parametric probability density function represented as a weighted sum of Gaussian component densities. Gaussian mixture models are commonly used as parametric models of the probability distribution of continuous measurements or features. Model-based clustering [18] is based on a finite mixture of distributions, in which each mixture component corresponds to a different cluster or block. For continuous data, the most common component distribution is a Gaussian distribution. Choosing a suitable number of components gc is essential for creating a useful model of the data and for data partitioning. The authors of [19] state that when a Gaussian mixture model is used for clustering, there might be an overestimation of the number of clusters. This is because a cluster may be better represented by a mixture of Gaussians than by a single Gaussian distribution. In [20] the authors argue that the goal of clustering is not the same as that of estimating the best approximating mixture model. Indeed, our objective in this work is not to find the number of components that best approximates the data, as an estimation for the number of blocks in each dimension's partition. Instead, we determine the minimum number of Gaussian components associated to each dimension's data, whose Root-Mean-Square deviation (RMSD) corresponds to a local minimum or approximates a minimum since the RMSD global minimum usually corresponds to a high number of components.

The process of choosing the number of components gc for the data associated to a dimension from a sample conformation data set is demonstrated in Figure 2. It starts with the generation of a fine histogram with N bins corresponding to a sequence of uniformly spaced single-valued points $\{x_k : k = 1,\ldots,N\}$ with associated data values $\{y_k : k = 1,\ldots,N\}$. Then, a set of m Gaussian models with a number of components varying from 1 to m is used for fitting the histogram's data. The Gaussian model is given by Eq. 1 where a corresponds to the amplitude, b is the centroid location, c is related to the peak width and m is the number of peaks to fit.

$$f(x) = \sum_{i=1}^{m} a_i e^{\left[-\left(\frac{x-b_i}{c_i}\right)^2\right]} \tag{1}$$

We analyze the curve generated for the graph where the x-axis represents the number of components $\{x_i : i = 1,\ldots,m\}$ and the y-axis corresponds to the associated RMSD values. We choose the smallest number of components with an RMSD corresponding to a local minimum or to a value that approximates a minimum.

Once the number of components corresponding to the number of clusters for each dimension is estimated, a Gaussian model-based clustering is used to partition the dimensions. Each element of the conformation's feature vector is assigned to a unique cluster in the corresponding dimension and this information is stored in B.



(a) (b) (c)

Fig. 2. Illustration of data modeling process: (a) histogram of data in a single dimension, (b) plot of RMSD versus number of Gaussian components, (c) data fitting with four Gaussian components

Once $\mathscr{L}$ and B are generated, the initial conformation clusters are created taking into consideration only the data and clusters from dimension $\mathscr{L}[1]$. Then, for each pair of dimensions $(\mathscr{L}[i], \mathscr{L}[i+1])$, we continue to refine our initial set of clusters by grouping the vectors that belong to the same cluster, i.e., those vectors that fall into the same block in $\mathscr{L}[i+1]$. The vectors comprising the cluster must also satisfy the constraint given by τ, whereby any "diverging" set of vectors must have a number of elements greater than a fraction τ of the total number of vectors in the original cluster. Vectors that do not satisfy this constraint are considered outliers or conformations that are unlikely to exist as meaningful intermediates; such vectors are removed from the clustering process. The final set of clusters is formed after all consecutive pairs of dimensions have been considered according to the order given by $\mathscr{L}$ and the process described above.

The algorithm for the proposed clustering method takes as input $\mathscr{C}$ and τ, and outputs a set of disjoint clusters as well as a set of outliers. The name of each final cluster shows the identification of the corresponding dimension blocks. The pseudocode is presented as Algorithm 1.

Algorithm 1 . Clustering Algorithm

```
Require: $\mathscr{C}$, $\tau$
Ensure: A set of disjoint clusters and a set of outliers
  Compute $\Sigma$
  Compute $\mathscr{L}$
  Compute $B$
  $i \leftarrow 1$
  Name each initial cluster in $\mathscr{L}[i]$ according to the block from which it originated
  for all consecutive pairs ($\mathscr{L}[i]$,$\mathscr{L}[i+1]$) do
    for all clusters up to dimension $\mathscr{L}[i]$ do Evaluate $T_\beta = \tau *$cardinality of cluster $\beta$
    end for
    Use $B$ to find and group the vectors that belong to same cluster $\beta$ up to dimension $\mathscr{L}[i]$ and
  fall into the same block at dimension $\mathscr{L}[i+1]$
    Obtain the cardinality $T_{g\beta}$ of each group $g$, where $\beta$ represents the original cluster from
  which those vectors originated
    if $T_{g\beta} \geq T_\beta$ then
      Discard $\beta$ and create a new cluster with the corresponding vectors
      Add the current dimension block identification (at dimension $\mathscr{L}[i+1]$) to the name of the
  original cluster $\beta$ in order to name the new cluster
    else
      Remove corresponding vectors from clustering process and classify them as outliers
    end if
  end for
```

The implementation of the algorithm includes MATLAB scripts and C code.

3 Experimental Results

3.1 Tested System - Adenylate Kinase (AdK)

The conformational search and subsequent clustering was run on AdK. It is a monomeric phosphotransferase enzyme that catalyzes reversible transfer of a phosphoryl group from ATP to AMP. The structure of AdK, which contains 214 amino acids, is composed of the three main domains, the CORE (residues 1–29, 68–117, and 161–214), the ATP binding domain called the LID (residues 118–167), and the NMP binding domain (residues 30–67). AdK assumes an "open" conformation in the unligated structure and a "closed"conformation. The RMSD between the two structures is 6.95Å. Supposedly, during the transition from the "open" to "closed" form, the largest conformational change occurs in the LID and NMP domain with the CORE domain being relatively rigid. Our model contains 8 rigid elements where most of the CORE domain was modeled as one large segment and was considered fixed, since it does not undergo a large-scale motion. Hence, the data were represented as feature vectors of 8 dimensions each. We ran the search 30 times in the direction of 1AKE-4AKE and 30 times in the reverse direction. Overall we collected 11,823 intermediate conformations.

3.2 Resulting Clusters

The data set $\mathscr{C}$ generated by our conformational search consists of the 60 pathways containing 11,823 conformations projected onto an 8-dimensional space which represents the protein conformational space. In order to perform a model-clustering of each dimension, the number of blocks in each partition is determined based on the number of components of the Gaussian mixture model chosen. The process for choosing the number of components generates histograms with 100 bins and analyzes Gaussian models with up to 8 components for each dimension. We use MATLAB with default arguments and the Trust-region algorithm [21, 22] to generate each model. We also used MATLAB to run the Gaussian model-based clustering, with the number of mixture components as input argument.

Our experiments use 4 different values of τ, the threshold used for identifying outliers: 0.05, 0.1, 0.2 and 0.3. Among all clusters generated by our method, we are interested in the large clusters (with at least 20 members), which distribute narrowly around their cluster center. While several of the most populated clusters may contain conformations that are close to known intermediates, some of them are narrower than others, regarding their deviation in terms of the centroid location. Since we are evaluating the RMSD between known structures and the resulting clusters centroid location, narrower clusters may produce more desirable results because their standard deviation is lower. We observed that for our sample data set, some clusters having at least 20 conformations contain conformations that are close to known intermediates structures (see Section 3.3). Further study is needed to determine the appropriate size for cluster of interest taking into consideration the standard deviation with respect to the centroid location and RMSD from known structures. Table 1 presents statistics about the resulting clusters according to selected values of τ. Notice that as the value of τ increases, so does the number of outliers detected by the algorithm.

3.3 Comparison with Known Intermediates

In general, knowledge about intermediate conformations is needed in order to provide a case-specific validation, but this knowledge does not always exist. As a matter of fact, intermediate structures are hard to obtain due to their relative high energy with respect to the native structures. With the advances in structural detection and simulation methods, one can expect to have more information about intermediate states in the future. AdK has several known mutant and intermediate structures. In a recent study [23] the energy profile of AdK was produced using elastic network interpolation (ENI).

Table 1. Statistics containing the number of resulting clusters (not considering the cluster containing the set of outliers), number of outliers and number of clusters with at least 20 conformations as well as the number of conformations in the smallest and largest clusters according to selected values of τ

	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$
n^o of clusters	329	231	58	10
n^o of outliers	567	1703	4380	7547
n^o of clusters with 20^+ conformations	88	75	44	10
size of smallest cluster	1	1	2	85
size of largest cluster	1312	1312	1312	1312

The method was used to generate the conformational transition pathway between the open and closed form of AdK and back, and compare the intermediates to known structural intermediates. Inspired by that study, we performed a similar test on our results. We focused on five known intermediates: chains A, B, and C of the hetero-trimer Adenylate Kinase from Aquifex Aeolicus (PDB accession code 2RH5), which are conformational change intermediates of the ligand free AdK [24], 1E4Y, which is an AdK mutant having 99% sequence identity with 4AKE and 1AKE and is a closed form of AdK binding with AP5, and 1DVR, which is a mutant that exhibits LID closure [25]. We selected all the clusters that containted at least 20 members and recorded for each cluster center the closest conformation to 1E4Y, 1DVR and to chains A, B and C of 2RH5. Our results are shown in Table 2. For each intermediate, the table shows the lowest RMSD from the closest conformation cluster center and the cluster number. We considered only "well-behaved" clusters, that is - the maximum distance from the cluster average was at most 3Å. Figure 4 shows the distribution of the RMSDs of cluster elements from the cluster center for these clusters. Notice that the same intermediate was the closest to chains B and C of 2RHC. This can be explained by the fact that the two chains are very similar to one another – the RMSD between them is approximately 2Å. The data corresponds to the result of our clustering method with a threshold τ of 0.05. In fact, this value of τ provided the best result obtained by our method. Figure 3 shows the intermediate structures superimposed on the closest cluster center for each intermediate.

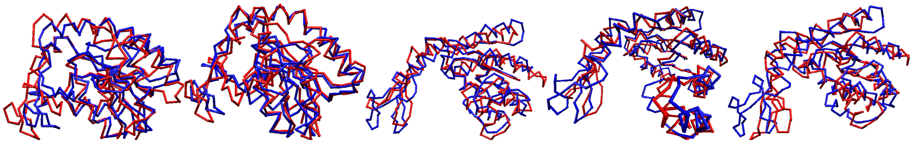
Table 2. RMSDs of cluster centers from five known AdK mutants representing intermediate states. The data were taken from our proposed method, cutoff of 0.05.

Intermediate PDB code	**2RH5(A)**	**2RH5(B)**	**2RH5(C)**	**1E4Y**	**1DVR**
Cluster name	1_2_3_1_1_1_2_3	1_2_1_2_1_1_2_3	1_2_1_2_1_1_2_3	1_2_2_2_2_1_3_3	1_1_3_2_2_1_2_1
Cluster size	33	21	21	84	101
RMSD with cluster average†	2.55	2.49	2.89	2.56	2.77

† The RMSD was calculated with respect to the $C - \alpha$ atoms of the aligned residues between the two proteins

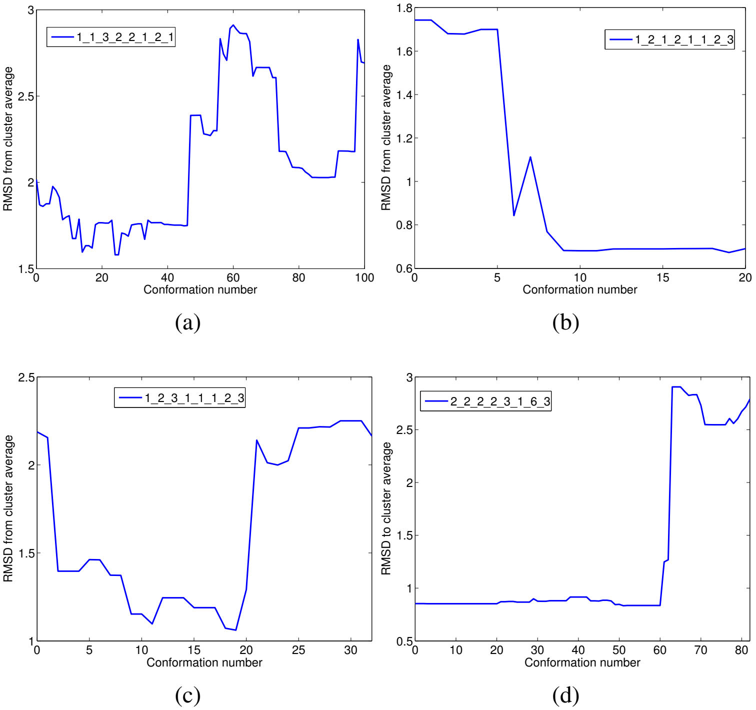
3.4 Comparison of the Proposed Clustering with K-Means

In order to validate our clustering algorithm, we compare our results to others generated by the K-Means algorithm. Weka [26] was the workbench used to run K-Means



(a) 1DVR (b) 1E4Y (c) 2RH5 (chain A) (d) 2RH5 (chain B) (e) 2RH5 (chain C)

Fig. 3. Cluster representatives (blue) superimposed on known intermediates (red). See Table 2 for details.



(a) (b) (c) (d)

Fig. 4. Distribution of RMSD of cluster elements from cluster average for the clusters representing AdK intermediates (see Figure 3 and Table 2. See inset legend for cluster name.

with Euclidian distance and a maximum of 2000 iterations. We selected 3 well spaced arbitrary values for the number of clusters K within the range of the number of clusters found by our method (see Table 1): 20, 80 and 150. Table 3 shows the results according to the number of clusters K. As seen, with $K = 20$ clusters there was no narrow cluster (with a radius below 3Å) corresponding to intermediates 1DVR and 1E4Y within a reasonable RMSD. Only at $K = 150$ the results were comparable to our method.

Table 3. RMSDs of cluster centers generated by K-Means from five known AdK mutants representing intermediate states

	2RH5(A)		**2RH5(B)**		**2RH5(C)**		**1E4Y**		**1DVR**	
K	RMSD	size	RMSD	size	RMSD	size	RMSD	size	RMSD	size
20	2.69	91	2.73	91	3.35	91	–	–	–	–
80	2.64	28	2.71	112	3.17	36	2.58	55	2.88	20
150	2.52	31	2.43	120	2.95	43	2.52	77	2.82	82

The proteins conformational space may contain many intermediate structures that are unlikely to exist as significant intermediates. Therefore, outlier detection is highly desirable in a clustering algorithm for the characterization of protein conformational space. The K-Means algorithm does not allow for the detection of outliers whereas our clustering method has the advantage of providing a flexible way to detect them. In addition, there is no a-priory way to know K, the number of clusters, in advance, and an educated guess has to be made. Our method provides a more deterministic way to evaluate the number of clusters. As a matter of fact, in this paper K was determined according to the number of clusters discovered by our method (see Table 1).

4 Conclusion

Characterization of protein conformational space is a very challenging problem due to the large amount of calculations required to characterize that complex and multi-dimensional space and due to the scarcity of experimental data regarding intermediate states. In this paper we presented a clustering method based on parallel coordinates and used it to characterize the conformational space of AdK and detect highly populated areas that may correspond to intermediate structures, which are usually hard to detect using experimental methods. In the case of AdK, however, several intermediate homologs exist and we were able to find cluster centers corresponding to these intermediates. The advantage of our method over K-means clustering and other standard clustering methods is that it allows the detection of outliers and does not require the number of final clusters to be given as input. Also, the parameters can be adjusted to gain insight about the optimal number of clusters. Detecting the ideal cutoff for the data and trying to find better ways to merge close clusters is the subject of on-going research.

References

1. Perutz, M.F.: Mechanisms of cooperativity and allosteric regulation in proteins. Quart. Rev. Biophys. 22, 139–236 (1989)
2. Case, D.A., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.: The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668–1688 (2005)
3. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
4. Haspel, N., Moll, M., Baker, M., Chiu, W., Kavraki, L.E.: Tracing conformational changes in proteins. BMC Structural Biology (2010) (in press)

5. Thomas, S., Tang, X., Tapia, L., Amato, N.M.: Simulating protein motions with rigidity analysis. J. Comp. Biol. 14(6), 839–855 (2007)
6. Chiang, T.H., Apaydin, M.S., Brutlag, D.L., Hsu, D., Latombe, J.-C.: Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics. J. Comp. Biol. 14(5), 578–593 (2007)
7. Raveh, B., Enosh, A., Furman-Schueler, O., Halperin, D.: Rapid sampling of molecular motions with prior information constraints. Plos Comp. Biol. (2009) (in press)
8. Zheng, W., Brooks, B.: Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. J. Mol. Biol. 346(3), 745–759 (2005)
9. Schroeder, G., Brunger, A.T., Levitt, M.: Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15, 1630–1641 (2007)
10. Weiss, D.R., Levitt, M.: Can morphing methods predict intermediate structures? J. Mol. Biol. 385, 665–674 (2009)
11. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall (1988)
12. McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–296 (1967)
13. Shehu, A., Kavraki, L.E., Clementi, C.: Multiscale characterization of protein conformational ensembles. Proteins: Structure, Function and Bioinformatics (2009)
14. Ladd, A.M.: Motion Planning for Physical Simulation. PhD thesis, Dept. of Computer Science, Rice University, Houston, TX (December 2006)
15. Inselberg, A.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: Proceedings of the First IEEE Conference on Visualization, California, USA, pp. 361–378 (1990)
16. McLachlan, G., Peel, D.: Finite Mixture Models. John Wiley and Sons (2000)
17. Inselberg, A.: Visual data mining with parallel coordinates. Computational Statistics 13 (1998)
18. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 611–631 (June 2002)
19. Baudry, J., Raftery, A.E., Celeux, G., Lo, K., Gottardo, R.: Combining mixture components for clustering. Journal of Computational and Graphical Statistics 19(2), 332–353 (2010)
20. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 719–725 (2000)
21. Celis, M.R., Dennis, J.E., Tapia, R.A.: A trust region strategy for nonlinear equality constrained optimization. In: Proceedings of the SIAM Conference on Numerical Optimization, pp. 71–82 (1984)
22. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust-Region Methods. SIAM, PA (2000)
23. Feng, Y., Yang, L., Kloczkowski, A., Jernigan, R.L.: The energy profiles of atomic conformational transition intermediates of adenylate kinase. Proteins 77(3), 551–558 (2009)
24. Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M., Hübner, C.G., Kern, D.: Intrinsic motions along an enzymatic reaction trajectory. Nature 450(7171), 838–844 (2007)
25. Schlauderer, G.J., Proba, K., Schulz, G.E.: Intrinsic motions along an enzymatic reaction trajectory. J. Mol. Biol. 256, 223–227 (1996)
26. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361 (1994)

A Method of Extracting Sentences Containing Protein Function Information from Articles by Iterative Learning with Feature Update

Kazunori Miyanishi[1] and Takenao Ohkawa[2]

[1] Graduate School of Science and Technology, Kobe University
1-1, Rokkodai, Nada, Kobe 657–8501 Japan
[2] Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe 657–8501 Japan
ohkawa@kobe-u.ac.jp

Abstract. Proteins are important macromolecules in living systems and serve various functions in almost all biological processes. Protein function information is reported in many scientific articles. Extraction of the function information from the articles is useful for drug discovery, understanding of life phenomenon, and so on. However, it is infeasible to extract the function information manually from a number of articles. In this paper, we propose a method of extracting sentences containing protein function information by iterative learning with feature update. In this method, we use a classifier in order to distinguish the sentences containing the function information from the other sentences, and introduce a semi-automatic procedure, in which a new classifier is reconstructed based on the user's feedback for the previous classified results. In the experiment with twelve articles as feedback data, it was confirmed that F-measure was improved by iterating learning without getting the negative effect of the feedback.

Keywords: protein function information, information extraction, decision tree, iterative learning.

1 Introduction

Protein serve various functions by interacting with other chemical compounds, and plays important roles in living systems[1]. Protein function is clarified by protein structure analysis and the obtained knowledge has been stated in a number of scientific articles. In order to make the knowledge available readily, it is required to construct a database storing protein function information. Many protein-related databases have been developed (e.g. PIR (Protein Information Resource)[2], PDB (Protein Data Bank)[3], Swiss-Prot (Swiss Protein Database)[4]). However, the useful information that has not been registered in such databases is still contained in huge volumes of articles.

Recently, there have been many attempts to extract significant information from biomedical documents. For example, Tsai et al.[5] and Sun et al.[6] proposed an approach to biomedical named entities recognition using Conditional

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 81–94, 2013.

Random Fields (CRF) [7] based on orthographical features or word conjunctions. And Seki et al. proposed a method using a protein name dictionary and rules for detection and filters[8]. On the other hand, researches to extract protein interactions information from biomedical literatures have been also conducted. Bunescu et al. attempted to identify human protein names and extract protein interactions comprehensively using various information extraction (IE) methods[9]; dictionary-based extraction, Rapier[10] and BWI[11] (a rule learning algorithm), Hidden Markov Models (HMMs)[12], Support Vector Machine (SVM)[13], and existing protein name identification systems (KEX[14] and ABGene[15]). Cooper et al. proposed a method for the discovery of protein-protein interactions using a combination of linguistic information (for example, verbs used to describe protein interactions) and graphical relations between proteins[16]. Hao et al. proposed an approach to discovering English expression patterns, optimizing them and extracting protein-protein interactions using them[17].

While these researches aim to extract biomedical named entities or protein interactions, we focus on literature about proteins whose structures are analyzed and registered in PDB, and have proposed a framework which assists the user in extracting protein function information interactively. In our scheme, a concept of extracting sentences containing the protein function information by iterative learning[18] has been introduced. Extraction of the sentences can be considered to classify sentences based on whether they contain the function information or not. In the previous method[18], the SVM (Support Vector Machine) is used as a classifier (we call this method 'SVM-based method'), where one sentence corresponds to one instance, and characteristics (keywords, patterns, etc.) of each sentence corresponds to the features of the instance. Each time receiving a feedback, which implies true/false evaluation for the classified sentences and new features automatically generated from these sentences, a whole classifier is reconstructed based on the training data including new features and new instances from the feedback, because in the SVM generally it is difficult to reconstruct just a part of the classifier that causes misclassifications. The method of incremental learning of SVM, where a new classifier is learnt not from scratch but by updating the current classifier every time a new instance is given, has been developed[19]. However, this method does not consider the feature update, namely the situation that new features (new attributes) as well as new instances are given incrementally. Although the feedback often improves classification accuracy in many cases, if the new classifier is built based on the feedback without referring the current classifier, the accuracy of the classifier may not be stable in iteration steps.

In this paper, we propose a new method of reconstructing a classifier partly based on a feedback in each iterative learning phase in order to make the accuracy rising steadily. As mentioned above, an SVM cannot flexibly respond to a feedback, because the inside of the model of an SVM is invisible. Therefore we use a decision tree[20][21] for the classifier. A decision tree consists of the combination of rules as a tree-structure. In the tree-structure, each node has one rule, each rule is expressed by features and each instance is classified according to whether it satisfies the condition of the rule at the node. It is more visible how instances are classified, compared with an SVM. There have been a lot of approaches to the incremental learning using the decision tree[22][23]. However, few of them suppose that new features are given by a feedback incrementally. In a decision tree, the feature at the node which frequently makes misclassifications can be easily identified. In our method, more accurate classifiers can be built by removing such wrong features and by adding new features and instances obtained from the user's feedback. In addition, by reconstructing not only a whole of decision tree but also a subset at which more misclassifications are taken place, a subset where instances are correctly classified can be kept in iteration. It is expected that the reconstructed decision tree is more accurate than the previous one.

2 Extracting Protein Function Information by Iterative Learning

In this section, the outline of iterative learning is shown firstly. Next features and the way to generate new patterns as one of the features are described. Then the algorithms of identifying the features which have a negative effect and of building a new classifier are described. Finally the whole procedure of the proposed method is summarized.

2.1 Outline of Iterative Learning

If enough instances and features as training data are not given in advance, an accurate classifier cannot be built based on only the training data. However, each time a classifier is applied to a new article, if the result of the classification is evaluated by a user, the classifier may be modified based on the user's feedback. Then it is expected to improve the accuracy of the classifier by iterating above processes.

We have proposed the SVM-based method of sentence classification using iterative learning in[18]. In this method, multiple features are assigned to each sentence, and a classifier is built based on them. In each iterative learning phase, a user determines whether the classified result is true or false. If false, patterns and instances extracted from the result are added to the training data, and a classifier is reconstructed and taken advantage of for next learning. Figure 1 shows the outline of the iterative learning, in which tagged document is used as the training data, and an SVM is used as a learner and a classifier.

On the other hand, in the new method proposed in this paper, a decision tree is used as a classifier. A decision tree consists of the combination of rules as a tree-structure in Figure 2. In a decision tree, each node corresponds to one rule and each rule is defined by one or more features. Instances are classified according to whether it satisfies a condition of the rule at the node. Therefore when the classified result is evaluated, it is possible to identify the features which cause misclassifications. In addition, it is possible to reuse the subset of

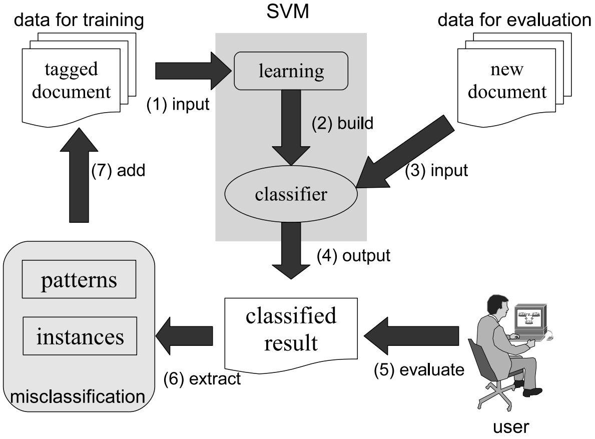


Fig. 1. The outline of the SVM-based method

a decision tree where instances are correctly classified because the inside of a decision tree is observable. In the proposed method, these characteristics are employed in iterative learning, and a classifier (a decision tree) is refined by removing features which have a negative effect. We call this method 'IDTFU (**I**terative **D**ecision **T**ree learning with **F**eature **U**pdate)'. The outline of the IDTFU is shown in Figure 3.

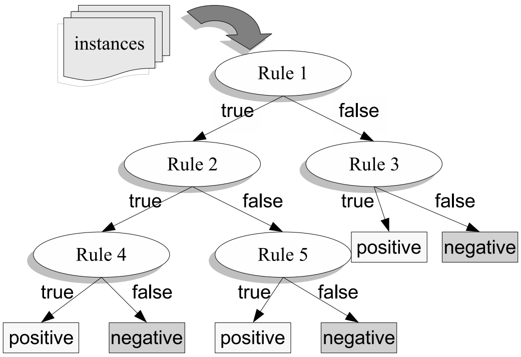


Fig. 2. Decision Tree

Unlike the SVM-based method in Figure 1, features with a negative effect are extracted from a classifier and removed at the steps (8) and (9) in Figure 3. And the new classifier can be built based on the subset of the previous classifier where instances are correctly classified.

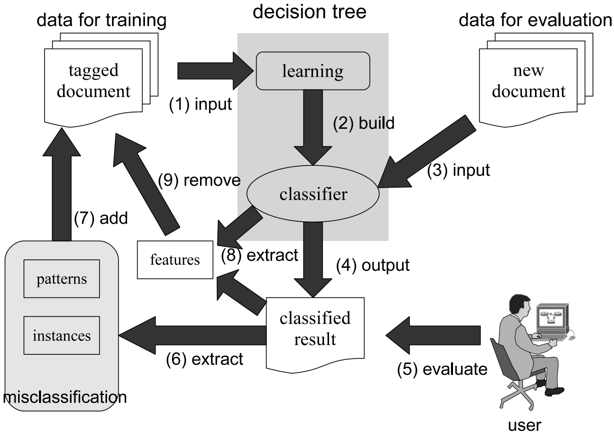


Fig. 3. The outline of the IDTFU

2.2 Features about Protein Function Information

Protein function infomation can be observed in articles about protein structure analysis. Sentences which contain protein function information have some characteristics. That is to say, the sentences may include names of substances and interactions, or sequences of some keywords. A classifier is built by using these characteristics as features for training data. The kinds of features used in the IDTFU are described below. Here, part-of-speech tag is put to each sentence in the documents using Brill's Tagger[24] in advance. And some types of named entities are tagged as each meaning (for example, <protein>, <residue>, <atom>, etc)[25].

1. Atomic distances between interacting substances
   When an amino acid residue interacts a substance, an atom or a part of atoms in the residue gets close to the substance. Therefore if the distance in three-dimensional space between a residue and a substance written in a sentence is shorter than a certain pre-defined threshold value, it is considered that the residue interacts the substanc[26] and "1" is assigned, otherwise "0" is assigned as a value of one feature for the sentence.
2. Keywords
   Frequently occurring words in sentences containing protein function information are significant as a hint for classification. Thus if the keywords, for example "interact", "bind", "hydrogen bond", and so on, are included in a sentence, "1" is assigned to the sentence, otherwise "0" is assigned.
3. Patterns
   Frequently occurring sequences of words in sentences conatining protein function information are also the hint for classification. These sequences are defined as patterns with wildcard characters, and used as features. For example, "<residue> (.)* play (.)* <function>", "<protein> (.)* contain

(.)* <residue>", where "<residue>" means the name of a residue, for example "Arg21" and "His23". Similarly, "<function>" and "<protein>" mean respectively the name of a function and a protein. If a sentence matches one of these patterns, "1" is assigned to the sentence, otherwise "0" is assigned.

2.3 Generation of New Patterns

In the IDTFU, patterns are generated automatically from misclassified instances and added to a set of training data as features in each iterative learning phase. If the target sentence includes words that are inappropriate for generating patterns, such words are removed, for example, a pronoun, definite or indefinite articles, and so on. In addition, words except a named entity, a verb and a noun are also removed.

For example, from the sentence *"At <residue> 270Glu1S </residue>, the electron density is again good and clearly shows <interaction> a stacked ring interaction </interaction> between <residue> 281Phe1L </residue> and <residue> 286Phe1F </residue>."*[27], the following six patterns are generated.

```
<residue>(.)*electron(.)*density(.)*<residue>
<residue>(.)*electron(.)*shows(.)*<residue>
<residue>(.)*electron(.)*<residue>
<residue>(.)*density(.)*shows(.)*<residue>
<residue>(.)*density(.)*<residue>
<residue>(.)*shows(.)*<residue>
```

2.4 Identification of the Features to Be Removed

The features monotonously increase by the pattern generation. If training data include features which disturb the improvement of the accuracy of the classifier, they should be removed. In the IDTFU, decision tree is used as a classifier and each node in the decision tree describes a condition of a feature. The feature of the node at which more instances are misclassified is the candiate to be removed. If the number of misclassified instances and the ratio of misclassification at a node are more than the thresholds T_{n_1} and T_{r_1} respectively, the path from the node to the root node is identified as the candidate to be removed (Figure 4).

The procedure of identifying the candidate to be removed is described in Figure 5, where L_{ALL} is a set of all leaves in the current decision tree, and C is the set of candidate leaves. NUM–MISS(l) returns the number of misclassification at the node l, and MISS–RATIO(l) returns the misclassification ratio at the node l.

Next, as shown in Figure 6, the node at the bottom of the candidate path is removed tentatively and the accuracy (F-measure) is obtained by Cross Validation (CV). If the accuracy obtained after removing the node is higher than one obtained before removing, the feature which the node indicates is determined to

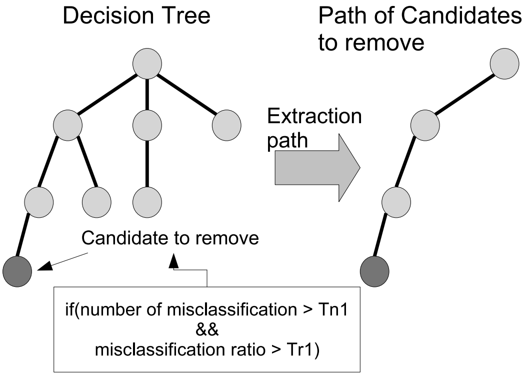

Decision Tree
Path of Candidates to remove
Extraction path
Candidate to remove
if(number of misclassification > Tn1
&&
misclassification ratio > Tr1)

Fig. 4. Extraction of the candidate to be removed

```
Procedure: cand_ident (Input : $L_{ALL}$ Output : $C$)
initialize : $C = \emptyset$
$\forall\, l \in L_{ALL}$
  $n$ := NUM-MISS$(l)$
  $r$ := MISS-RATIO$(l)$
  if$(n > T_{n1}$ && $r > T_{r1})$
      $C := C \cup l$
return $C$
```

Fig. 5. Procedure: *cand_ident* for identifying a candidate node to be removed

be removed. In addition, the parent node of the removed node is treated as a new candidate to be removed. This process is continued while the accuracy is improving.

This procedure is formally described in Figure 7, where N_c is each candidate leaf identified by the procedure *cand_ident*, and N_r is the node to be removed. T_c is the current decision tree, and CALC–F(T) returns F-measure of the decision tree T. F_t is F-measure of the current decision tree by Cross Validation and used as the threshold. REMOVE–PARENT(N, T) removes the parent node of N from T, and PARENT(N) returns the parent node of N.

2.5 Reconstruction of Decision Tree

The simplest way to update a decision tree is reconstruction of a whole decision tree. When a current learner gets user's feedback, new features and new instances are added to current training data and some features are removed from them based on the feedback. Then, new decision tree can be built from scratch using the new training data. The IDTFU provides the following more effective strategies to reconstruct a decision tree.

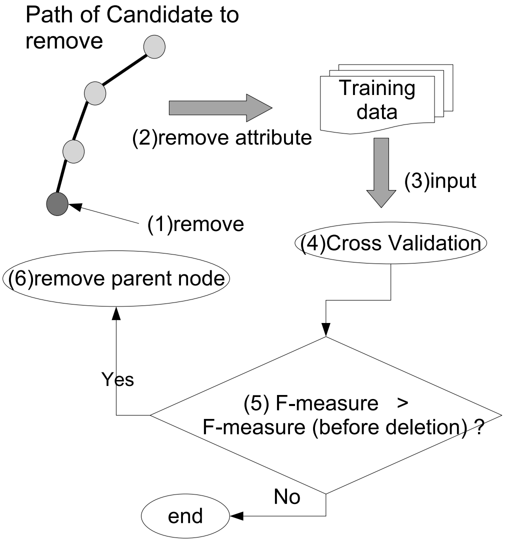


Fig. 6. Determination of the node to be removed

```
  Procedure: $node\_det$ (Input : $N_c$ Output : $N_r$)
1 $F_t$ := CALC−F($T_c$)
2 $N_r \leftarrow$ NULL
3 $N_p \leftarrow N_c$
4 $T_r$ := REMOVE−PARENT($N_p, T_c$)
5 $F_c$ := CALC−F($T_r$)
6 if ($F_c > F_t$)
      $N_r$ := $N_p$
      $N_p$ := PARENT($N_p$)
      go to 4
7 else
      return $N_r$
```

Fig. 7. Procedure: *node_det* for determining a node to be removed

(a) Add Features, Remove Features and Reconstruct a Subset of the Decision Tree. The first strategy is to reconstruct a subset of decision tree. It may not always be efficient to reconstruct a whole decision tree every time getting user's feedback. In this case, only the subset which frequently misclassifies should be reconstructed. After the node is removed by the method mentioned in 2.4, the edge from the node is reconstructed as shown in Figure 8.

(b) Same as strategy (a) If the Size of the Decision Tree is Small, Otherwise Reconstruct a Whole Decision Tree. The second strategy is a hybrid method. That is to say, a subset of decision tree is reconstructed while the

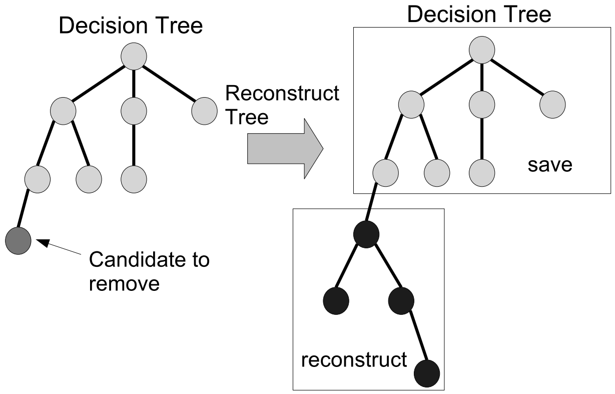


Fig. 8. Reconstruction of subset trees

size of decision tree is smaller than a certain threshold, and otherwise a whole decision tree is reconstructed. While a subset of decision tree is reconstructed in iterative learning, the number of misclassification at each node which is not reconstructed does not change. Therefore as the decision tree becomes larger, the sum of misclassification may be increasing. In addition, it is generally preferred that the size of decision tree is small because the decision tree is unlikely to overfit to the training data. For these reasons, the hybrid method expects to be effective.

The procedure of reconstruction of subset trees shown in Figure 8 is described in Figure 9, where T_c is the current decision tree, N_r is the node to be removed based on the procedure *node_det*, and T_n is the reconstructed decision tree. $I(N)$ is a set of instances at the node N, $\textsc{remove}(T, N)$ removes the node N from the decision tree T. $\textsc{build-subtree}(T, I)$ builds a subtree based on instances I, add the subtree to the tree T, and returns a reconstructed tree.

```
Procedure: sub_reconst (Input : $T_c$, Output : $T_n$)
$T_r := \textsc{remove}(T_c, N_r)$
$T_n := \textsc{build-subtree}(T_r, I(N_r))$
```

Fig. 9. Procedure: *sub_reconst* for reconstructing subtree

The common part of both methods (a) and (b) is shown in Figure 10, where $\textsc{build}(D)$ builds a decision tree based on the training data D, and F_f is a set of new features obtained by the feedback. In this part, the first classifier is built based on the initial training data, and new features are added to the training data after receiving a feedback.

The procedure of the method (a) is shown in Figure 11. In the method (a), features which have a negative effect are removed from the current training data

```
   Procedure: common part of tree_reconst
1  $D \leftarrow$ initial data for training
2  $T := \textsc{build}(D)$
3  if receive the feedback
4      $D := D \cup F_f$
```

Fig. 10. The common part of procedure of tree reconstruction

at line **5**, where F_r is the features to be removed based on the procedure *node_det*, and $\textsc{remove-f}(D, F)$ removes features F from the data D. And the decision tree is partly reconstructed at line **6**.

The method (b) is shown in Figure 12, where T_s is the threshold value of the size of a decision tree. That is to say, while the size of the decision tree is smaller than T_s, the tree is reconstructed partly. If the size is larger than T_s, the tree is reconstructed whole.

```
     Procedure: tree_reconst_a
1-4  common part
5    $D := \textsc{remove-f}(D, F_r)$
6    $T := sub\_reconst(T)$
```

Fig. 11. The procedure of tree reconstruction (a)

```
     Procedure: tree_reconst_b
1-4  common part
5    if the size of $T < T_s$
6        $D := \textsc{remove-f}(D, F_r)$
7        go to 2
8    else
9        $T := \textsc{reconstruct}(T)$
```

Fig. 12. The procedure of tree reconstruction (b)

3 Evaluation

We evaluate the effectiveness of the IDTFU by using articles stating protein structural analysis shown in Table 1, each of which is referred by PDB. PDB-ID is the identifier of the protein registered in PDB, and the "correct sentence" means the sentence containing protein function information.

Named entities in these articles are already tagged manually. In our experiment, one article is used for training, another is for evaluation, and the others are used as feedback data. We conduct sixteen trials changing the combination

Table 1. Statistics of the articles referred in the experiment

PDB-ID	# of sentences	# of correct sentences
1a0h	289	26
1a0q	259	23
1a26	203	13
1a3l	214	23
1a3r	299	21
1a4j	190	13
1a5a	113	10
1a5h	245	39
1a5i	275	73
1a5v	241	20
1a5y	256	33
1a5z	304	8
2a2g	288	13
2a39	312	4

Table 2. The difference of feature update and reconstructing tree in each method

method	features	features	tree reconstruction	
	addition	removal	subset	whole
SVM-based	○			
IDTFU (a)	○	○	○	
IDTFU (b)	○	○	○	○

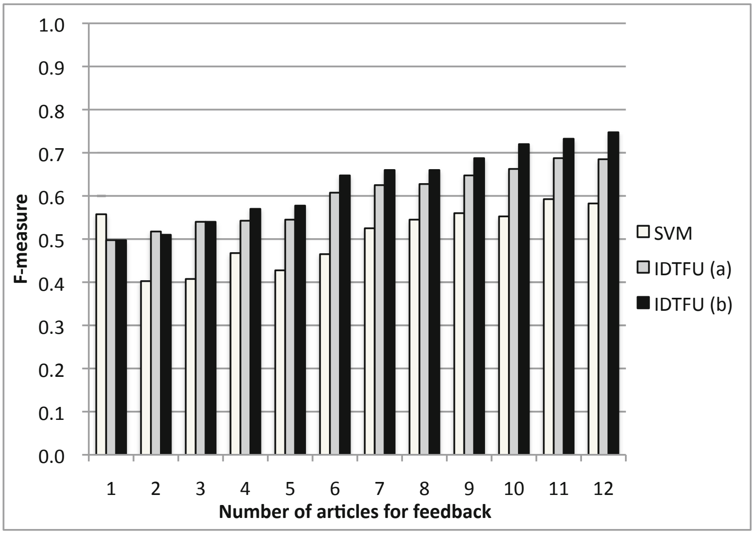


Fig. 13. Extraction accuracy in each method

of training data, evaluation data and feedback data. The order of feedback data is decided randomly.

We compare the IDTFU ((a) and (b)) with the SVM-based method in order to evaluate the effectiveness of the IDTFU. The outline about feature update or reconstruction of decision tree in each method is shown in Table 2. T_s, the threshold of the size of the decision tree, in Figure 12 is 40 including leaf nodes, because the size of the tree often becomes between 40 and 50 nodes after several iteration steps.

The average F-measure of each method for 16 trials is shown in Figure 13. Although the F-measure of the SVM-based method is the highest in the initial learning phase, the F-measure goes up and down depending on the feedback. On the other hand, the F-measure of the IDTFU rises monotonously by iterative learning and converges in higher value than the highest value obtained by the SVM-method.

From this experiment, it is obvious the accuracy of the IDTFU steadily rises by iterative learning based on the feedback. In addition, the IDTFU always has a certain level of accuracy when receiving any feedback, because the variance is small at each iteration step. In the IDTFU (b), the whole of the decision tree is reconstructed when the size of the tree becomes larger than a certain threshold. Therefore the classifier is reconstructed and often gets a higher accuracy than the accuracy that the IDTFU (a) shows.

4 Conclusion

In this paper, we proposed a method of extracting sentences containing protein function information by iterative learning with feature update. The IDTFU (a) is to update features and reconstruct the subset of the decision tree, and the IDTFU (b) is to update features and reconstruct the subset or the whole of the decision tree. We compared these methods with the SVM-based method, and evaluated the effectiveness of the IDTFU. While the accuracy of the SVM-based method rises and falls depending on the feedback and the variance of F-measure for the combination of training data, evaluation data and feedback data at each iteration step is large, the accuracy of the IDTFU usually rises steadily and the variance is small. In addition, we confirmed that the hybrid tree reconstruction, namely applying subset tree reconstruction or whole tree reconstruction selectively, was effective for improving accuracy of the iterative learning.

References

1. Berg, J., Tymoczko, J., Stryer, L.: Biochemistry, 5th edn., vol. 423, pp. 436–437. WH Freeman and Company (2002)
2. Wu, C.H., Yeh, L.S.L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., et al.: The protein information resource. Nucleic Acids Research 31, 345–347 (2003)

3. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Research 28, 235–242 (2000)
4. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al.: The swiss-prot protein knowledgebase and its supplement trembl in 2003. Nucleic Acids Research 31, 365–370 (2003)
5. Tsai, R.T.H., Sung, C.L., Dai, H.J., Hung, H.C., Sung, T.Y., Hsu, W.L.: Nerbio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC Bioinformatics 7(suppl. 5), S11 (2006)
6. Sun, C., Guan, Y., Wang, X., Lin, L.: Biomedical Named Entities Recognition Using Conditional Random Fields Model. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 1279–1288. Springer, Heidelberg (2006)
7. Lafferty, J., Pereira, F., McCallum, A.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, ICML 2001 (2001)
8. Seki, K., Mostafa, J.: An approach to protein name extraction using heuristics and a dictionary. In: The American Society for Information Science and Technology (ASIST) Annual Meeting, vol. 40, pp. 71–77 (2003)
9. Bunescu, R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K., Wong, Y.W.: Learning to extract proteins and their interactions from medline abstracts. In: Proceedings of the International Conference on Machine Learning 2003 Workshop on Machine Learning in Bioinformatics, pp. 46–53 (2003)
10. Califf, M.E., Mooney, R.J.: Relational learning of pattern-match rules for information extraction. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI 1999), pp. 328–334 (1999)
11. Freitag, D., Kushmerick, N.: Boosted wrapper induction. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 577–583 (2000)
12. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
13. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer (1995)
14. Fukuda, K., Tsunoda, T., Tamura, A., Takagi, T.: Information extraction: Identifying protein names from biological papers. In: Proceedings of the Pacific Symposium on Biocomputing, pp. 707–718 (1998)
15. Tanabe, L., Wilbur, W.J.: Tagging gene and protein names in biomedical text. Bioinformatics 18(8), 1124–1132 (2002)
16. Cooper, J.W., Kershenbaum, A.: Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. BMC Bioinformatics 6, 143 (2005)
17. Hao, Y., Zhu, X., Huang, M., Li, M.: Discovering patterns to extract protein-protein interactions from the literature: part ii. Bioinformatics 21(15), 3294–3300 (2005)
18. Munna, M.A., Ohkawa, T.: A method to extract sentences with protein functional information from literature by iterative learning of the corpus. IPSJ Transactions on Bioinformatics 47(SIG 17(TBIO 1)), 22–30 (2006)
19. Cauwenberghs, G., Poggio, T.: Incremental and decremental support vector machine learning. In: Proceedings of the Neural Information Processing Systems (NIPS 2000), vol. 13 (2001)

20. Quilan, J.R.: Decision trees and multi-valued attributes. Machine Intelligence 11, 305–318 (1988)
21. Quilan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
22. Utgoff, P.E.: Incremental induction of decision trees. Machine Learning 4, 161–186 (1989)
23. Domingos, P., Hulten, G.: Mining high-speed data streams. In: Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, pp. 71–80 (2000)
24. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics 21, 543–565 (1995)
25. Numa, M., Kaneta, Y., Ohkawa, T.: Automatic classification of proper names in protein-related literatures using database retrieval on www. In: Proceedings of the Fifth International Conference on Computational Biology and Genome Informatics, CBGI 2003, pp. 903–906 (2003)
26. Kaneta, Y., Munna, M.A., Ohkawa, T.: A method for extracting sentences related to protein interaction from literature using a structure database. In: Proceedings of the Second Workshop on Data Mining and Text Mining for Bioinformatics (in conjunction with ECML/PKDD 2004), pp. 18–25 (2004)
27. Martin, P.D., Malkowski, M.G., Box, J., Esmon, C.T., Edwards, B.F.P.: New insights into the regulation of the blood clotting cascade derived from the x-ray crystal structure of bovine meizothrombin des f1 in complex with ppack. Structure 5, 1681–1693 (1997)

Non-coding RNA Covariance Model Combination Using Mixed Primary-Secondary Structure Alignment

Jennifer A. Smith

Boise State University, Department of Electrical and Computer Engineering, Boise, Idaho, USA
jasmith@boisestate.edu

Abstract. Covariance models are very effective for finding new members of non-coding RNA sequence families in genomic data. However, the computation burden of applying CM-based search algorithms can be prohibitive. When annotating the genome of a newly sequenced organism it is usually desired to search the sequence data using a large number of ncRNA families. Computational burden can be reduced if the families are clustered into statistically similar models and a single cluster-average representative model produced. The database is then searched with the representative model for each cluster at a relatively low detection threshold. The output of this pre-filtered database is then processed with the individual family members of the cluster. A base-pair conflict metric has previously been proposed for use in model clustering. In this work an alternative metric using standard alignment algorithms and a special mixed primary-secondary structure scoring matrix is proposed.

Keywords: non-coding RNA, covariance model, sequence analysis, secondary structure.

1 Introduction

Reduction of the large computational burden imposed by covariance-model-based non-coding RNA (ncRNA) gene finding has been the subject of much research. Methods examined have included direct hardware acceleration [1], various database prefiltering operations and search space limiters [2-5], model simplifications [6], and generic (non-family specific) ncRNA search algorithms [7]. These have been applied with varying degrees of success and often the methods are not mutually exclusive and can be combined.

Covariance models [8-9] are preferred for ncRNA over primary-structure-only methods such as profile hidden Markov models due to the high degree of secondary structure conservation and low degree of primary structure conservation in these sequences. Intra-molecular bonding patterns map very well to three-dimensional shape and function in ncRNAs, whereas primary structure is a poor predictor of secondary structure. Portions of genomic sequence which code for protein can be converted to equivalent amino acid sequences with much more primary structure similarity within a protein family. A major cost associated with allowing both primary and secondary

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 95–104, 2013.

structure to be expressed in statistical models such as covariance models (CMs) is a large increase in database search computational effort. As compared to profile hidden Markov models [10], this increase can be as much as an order of magnitude and compared with faster algorithms such as BLAST [11-12] the increase may be many orders of magnitude.

A very popular strategy for ncRNA gene search has been to group known ncRNA genes into families and form a statistical model of each family using a multiple alignment of the family sequences mapped onto a secondary structure specification. Such a methodology may be found in the Rfam database [13-15], which uses the Infernal [16-17] suite of CM-based algorithms for model building and database search. The choice of how much variation to allow within a family versus breaking a family up into independent families is rather arbitrary and this choice has so far been treated as an art. Recent versions of the Rfam database have also included groups of families called clans. However, these clans are intended more for the purpose of presenting potential similarity in biological function than as a way to group families for improvement of efficiency in database search. Clans are useful as a sanity check on clustering used in model combination for computational efficiency purposes since we expect that families that are members of the same clan due to functional similarity are more likely to have structural similarity than families that are not members of the same clan.

Clustering of models can be taken to the extreme of insisting on placing all families in a single cluster and finding a single generic search model that can be used as a database pre-filter. This filter separates the database into two parts: sections of the database that do not seem to contain any ncRNA and those that possibly do contain one or more ncRNA genes. There is rather convincing evidence [18] that merely searching for database sections that have the potential to form stem-loop structures is not informative enough to do any significant amount of database reduction. There is a lot of diversity in the number of stem-loop structures and sizes of stems and loops between families as well as very little primary sequence similarity between families. However, there is generally a lot of similarity in secondary structure and a significant if not large amount of primary sequence similarity within families. This leads to modeling of families individually. The idea of clustering and combining models can be viewed as an intermediate between these two extremes.

Jiang and Wiese [19] have taken this clustering and model combining approach to search time reduction. In order to do the clustering, it is necessary to have a distance metric for comparing the models to be clustered. Their approach is to use a base pair conflict metric. This is a pair-wise metric that uses the dot-bracket RNA secondary structure notation for the two models as input and looks for instances where mapping one model onto the other would imply a pseudoknot. Covariance models do not allow pseudoknots and where they do occur in actual ncRNA families, they are simply ignored and treated as non-based-paired sequence locations. Combination of two models such that there is an implied pseudoknot is an indication that the two models are not amenable to such combination. Formally, if positions (m, n) are base paired in one model and positions (i, j) are based paired in another, then if m < i < n < j or i < m < j < n, there is an implied pseudoknot, or in the Jiang-Wiese terminology a base pair conflict. Note that the two base pairs (m, n) and (i, j) should be specified with m < n and i < j for this definition to work. Their distance metric is based on a count of the number of base pair conflicts.

The approach taken in this work is similar to that of Jiang and Wiese with the exception that an entirely different distance metric is proposed. The new metric uses the alignment score output of standard pair-wise or multiple alignment algorithms. A special 12-character alphabet for mixed primary-secondary structure description is used and an associated special-purpose scoring matrix is introduced. Since there are twelve characters in the alphabet and the characters are judiciously chosen to match valid amino acid symbols (of which there are 20), any alignment algorithm designed to handle proteins should work. As a side effect of scoring the alignment for a cluster with three or more family members using a multiple alignment algorithm, a consensus secondary structure is normally obtained. This consensus secondary structure may be used as a guide on which to build a combined covariance model. Furthermore, the metric uses both primary and secondary structure information, whereas the base pair conflict metric uses only secondary structure information.

The remainder of this work is structured as follows. Section 2 gives details of the special alphabet and scoring matrix for this application as well as the support-vector-machine clustering and combined model building methods. Section 3 compares the use of the proposed metric with that of the base pair conflict metric using the same data found in the Jiang and Wiese paper. Section 4 expands the analysis to a case of members of two Rfam clans with the result that the clustering tends to automatically place members of the same clan in the same clusters. The concluding section discusses work yet to be done as well as summarizing the results of the analysis sections.

2 Distance Metric, Clustering, and Model Combination

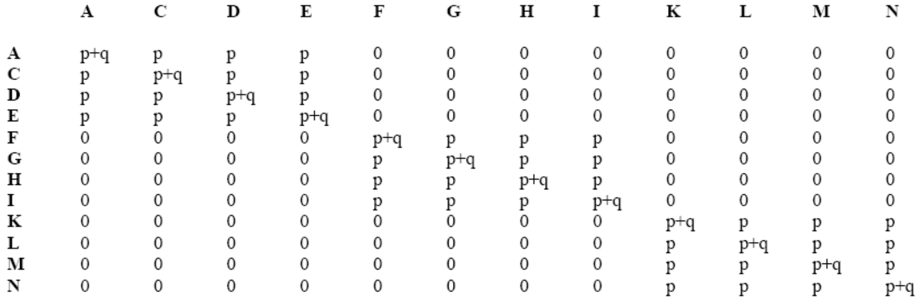
Three steps are required to obtain a reduced set of combined models for comprehensive annotation of a genome with ncRNA putative genes. First, a set of distances between family models for which searches are to be performed needs to be obtained. Knowledge about the type of organism the genome belongs to will likely lead to elimination of some of the Rfam families from consideration. For example, if the organism is a mammal and a certain family is known to only exist in bacteria, then it probably does not make sense to even consider that family. Second, these distances are used to cluster the family models. There is a tradeoff between fewer clusters resulting in faster processing and more clusters resulting in better search results. Third, the models in each cluster need to be combined into a single covariance model for use in the first round of search. A subsequent round of search will be applied to the results of each first round search using the original models for each of the members of the cluster.

2.1 Alphabet and Scoring Matrix

Stockholm format alignment files are available for each Rfam family which include "#=GC SS_cons" and "#=GC RF" lines giving consensus secondary structure and primary structure respectively. A script is written to extract these lines from each family of interest and a mixed primary-secondary structure symbol is generated for each alignment column. The 12-character alphabet for these symbols is shown in Figure 1a. The symbols are chosen to be valid amino acid symbols such that any standard alignment program that can handle protein sequences will not do anything unusual based on the assumption that positions are degenerate. All combinations of the four nucleotide possibilities (a, c, g and u) and the three secondary structure labels (dot, left bracket and right bracket) result in twelve possibilities. The easiest thing to do with any column that has a degenerate primary or secondary structure is simply to discard that column.

A:	a / .	**C**:	c / .	**D**:	g / .	**E**:	u / .
F:	a / [	**G**:	c / [	**H**:	g / [	**I**:	u / [
K:	a /]	**L**:	c /]	**M**:	g /]	**N**:	u /]

(a)



	A	C	D	E	F	G	H	I	K	L	M	N
A	p+q	p	p	p	0	0	0	0	0	0	0	0
C	p	p+q	p	p	0	0	0	0	0	0	0	0
D	p	p	p+q	p	0	0	0	0	0	0	0	0
E	p	p	p	p+q	0	0	0	0	0	0	0	0
F	0	0	0	0	p+q	p	p	p	0	0	0	0
G	0	0	0	0	p	p+q	p	p	0	0	0	0
H	0	0	0	0	p	p	p+q	p	0	0	0	0
I	0	0	0	0	p	p	p	p+q	0	0	0	0
K	0	0	0	0	0	0	0	0	p+q	p	p	p
L	0	0	0	0	0	0	0	0	p	p+q	p	p
M	0	0	0	0	0	0	0	0	p	p	p+q	p
N	0	0	0	0	0	0	0	0	p	p	p	p+q

(b)

Fig. 1. a) Mixed primary-secondary structure alphabet for use in standard alignment algorithms b) Special-purpose scoring matrix for distance metric calculation

Any protein alignment algorithm is then run using a scoring matrix of the form shown in Figure 1b. The value p is a positive quantity giving credit for matching the secondary structure and the value q is a positive quantity giving credit for matching the primary structure. The secondary structure is given in dot-bracket notation, where a dot indicates that the column is not base paired, a left bracket indicates that the column is intramolecularly paired with another column to the right and a right bracket indicates that the column is paired with another column to the left. Since pseudoknots are not allowed in covariance models, this notation is unambiguous as to which column is paired with which.

Since the number of families may be quite large, it is recommended not to generate a distance score for every pair of families. Instead, choose a small number of reference families at random (or perhaps, with some inspection, judiciously) and find distances to all families relative to these few. A pseudo-transitive estimate of family distances can then be found as a combination of the distances from one family to each reference family and from the other family to each reference family. We may similarly think of the set of distances from a family to each reference family as a feature set and use any clustering algorithm that can handle multiple features to form the families into clusters.

2.2 Support Vector Machine Clustering Method

The particular clustering method used for the results in this work is a support vector machine (SVM). SVM clustering is discussed in [20] and a Matlab toolbox to implement SVM clustering is available at [21]. A set of ten reference families were used such that a ten dimensional clustering was undertaken. In order to keep as much diversity in the reference families as possible, the following method was used. First, fifty families were chosen at random and the distances computed between all pairs. Then the two most distant families of the fifty were selected. A third was selected such that the average distance between the first two and the third was maximized. This pattern was repeated until a set of ten (hopefully very diverse) families were found.

2.3 Model Combination

In the paper by Jiang and Wiese [19], the method of model combination is a pair-wise combination process which selects base-paired alignment columns from one family or the other and unpaired alignment columns from both such that column numbers in the combined model exactly match column numbers in the original two family alignments. It appears that a combined alignment is generated and input to the cmbuild program of the Infernal program suite. Presumably, the combined alignment has number of sequences equal to the lesser of the number of sequences in each original family and that one half the number of combined alignment sequences is randomly chosen from each original alignment for unpaired columns. The output parameter file from the cmbuild program is then used in the cmsearch program for database search.

In this work another approach is used. Instead of building up combined models for more than two families by iterative pair-wise merging, it is done in a single iteration. First, all of the mixed primary-secondary structure alphabet representations of the families in a cluster are processed using a multiple alignment program. All of the sequences from all of the family members in the cluster are then placed in one big alignment file where the output of the multiple alignment on the mixed primary-secondary structure is used as a guide to placement in the combined alignment file. This can result insertions and deletions in individual sequences. Deleted columns simply do not appear in the final alignment. Inserted columns result in a unknown nucleotide, but there is a symbol available for such an unknown. The result is input in the cmbuild program to generate a parameter file. This method has the advantages of using all of the available sequence data to build the combined model and the use of multiple alignment to jointly find the secondary structure rather than building it up pair-wise sequentially, which generates a result that depends on the order of the build.

3 Comparision with Base Pair Conflict Measurement

The families used in the Jiang and Wiese [19] paper are taken from Rfam and have accession number RF00007, RF00020, RF00015, RF00029, and RF 00050. Respectively, these are U12 (minor spliceosomal RNA, 63 sequences), U5 (U5 spliceosomal RNA, 181 sequences), U4 (U4 spliceosomal RNA, 178 sequences), Intron_gpII (group II catalytic intron, 98 sequences), and FMN (FMN riboswitch -- RFN element, 146 sequences). The number of sequences listed is the number of 'seed' sequences, which are the sets of highly curated sequences used to build Rfam models. Far more putative members of these families exist in the Rfam database which are the result of search with the models. Hierarchical clustering of these five families resulted in the following grouping: {(U12 U5) [U4 (Intron_grpII FMN)]}.

Using the distance metric proposed in this work and SVM clustering into three groups results in grouping U12 and U5 together and grouping Intron_grpII and FMN together, which is consistent with the base pair conflict metric. However, using two groups results in grouping U12, U5, and U4 together and grouping Intron_grpII and FMN together, which is different than the base pair conflict results. In other words, the mixed primary-secondary alignment results in the grouping: {[(U12 U5) U4] (Intron_grpII FMN)}. Interestingly, the functionally similar spliceosomal families are grouped together.

Due to the small number of families studied in Jiang and Wiese [19] the advantages of using all the available sequence data and doing joint rather than pair-wise sequential structural combination is hard to access. For the five families chosen, the difference between the least and most number of sequences is not great (63 versus 181) and therefore the amount of information lost is relatively small. The U12/U5 combined model is based on 63 sequences, the Intron_gpII/FMN combined model is based on 98 sequences, and the U4/Intron_gpII/FMN combined model is based on 98 sequences (a five-way combined model is not reported). If one of the families had only four seed sequences (not an unusual situation in Rfam), the loss of information would be much greater. The fact that only the U4/Intron_gpII/FMN combined model has more than two families means that the use of multiple alignment in this work versus sequential pair-wise is of little significance.

Table 1 shows a comparison between the scores generated by the original Rfam models, the Jiang/Wiese combined models and combined models using the methods of the present work. Three sequences are analyzed from each family: AANN01056468, AAVX01293999, AAB01008960 (U12); CAAE01011861, Z149914, AATU010003637 (U5); ABDC013119198, AABS01000042, X67145 (U4); AJ315331, X55026, X04465 (Intron_gpII), CP000724, AE00633, L0922288 (FMN). Hereafter, these sequences are referred to as U12.1, U12.2, U12.3, U5.1, etc. The scores shown are log likelihood ratios (with the logs in base 2), so an increase of 1.0 in score is associated with a factor of two reduction in false alarm rate for a given sensitivity. Both the Jiang/Wiese results and the present work increase in false alarm rate over using individual models by a large amount.

Scores in the 40 to 50 range are generally considered to be acceptable for CM-based ncRNA gene search, so even though there is a very large degradation in score, many of the models in the tables may still be considered acceptable. In general, the scores using the methods of this work (MPS) are higher than those using the base pair conflict metric

Table 1. Original, Base Pair Conflict and Mixed Primary-Secondary Structure Metric Scores

	Original	Base Pair Conflict		Mixed Primary-Secondary Structure		
Sequence	No Grouping	U2 + U5 or gpII + FMN	U4 + gpII + FMN	U2 + U5 or gpII + FMN	U4 + gpII + FMN	U2 + U5 + U4
U12.1	155	35	-	72	-	64
U12.2	129	63	-	68	-	58
U12.3	119	40	-	45	-	39
U5.1	112	88	-	93	-	83
U5.2	95	72	-	81	-	75
U5.3	93	76	-	83	-	77
U4.1	125	-	19	-	35	72
U4.2	115	-	18	-	34	68
U4.3	104	-	3	-	28	61
gpII.1	64	58	17	53	23	-
gpII.2	63	14	12	52	27	-
gpII.3	53	11	14	48	22	-
FMN.1	120	71	11	88	43	-
FMN.2	118	71	13	89	41	-
FMN.2	94	58	2	64	32	-

(BPC). Two combined models for three families are shown for MPS. The combined model for U4/Intron_gpII/FMN would not normally be selected using the MPS methods since the clustering does not select this as a group, but is included for comparison with the BPC method. The combined model using MPS does in fact do better than that of BPC, but its usefulness is marginal, with scores in the 20-40 range. What is really striking is that the combined model for U12/U5/U4, which is the grouping MPS selects is much better. In fact, using the two combined models U12/U5/U4 and Intron_gpII/FMN from the MPS method covers all five families with no score less than 39.

Clearly a much larger study is needed to determine if the reduction in number of family models is large over the whole Rfam database, but these results do seem promising.

4 Results with Two Known RFAM Clans

As a further study of how the proposed distance metric performs, two clans were selected from the Rfam database: CL00014 and CL00015. These two clans were selected because they each contain several families. CL00014, the CRISPR-1 clan, contains the families CRISPR-DR2, CRISPR-DR4, CRISPR-DR14, CRISPR-DR17, CRISPR-DR25, CRISPR-DR43 and CRISPR-DR66. CL00015, the CRISPR-2 clan, contains the families CRISPR-DR5, CRISPR-DR7, CRISPR-DR63 and CRISPR-DR64. Many of the 102 clans listed in the Rfam database contain only two families. Also, these two clans are clearly related in that both contain Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), but the curators of Rfam believe that they are not so closely related as to warrant being in the same clan.

The best clustering results for the eleven families listed above seem to come from having four clusters. The first cluster has six members, four from the CRISPR-1clan : DR-66, DR-17, DR-25 and DR-4 and two from the CRISPR-2 clan: DR-5 and DR-63. The second cluster contains DR-64 and DR-7, both from the CRISPR-2 clan. The third cluster contains DR-14 and DR-43, both from the CRISPR-1 clan. The fourth cluster contains only DR-2 from the CRISPR-1 clan. Only one of the clusters contains members from two different clans. However, the clans organization is designed to group together ncRNA families that are functionally related, not families with similar sequences. The two are often consistent, but not always. A closer look at the structure of the families in the clusters is shown in Table 2, where the assigned clustering makes a lot more sense. Cluster 1 contains a single stem-loop structure with a short loop and 5 to 7 base pairs in the stem. Cluster 2 contains a single stem-loop with a longer loop. Cluster 3 contains no base pairing and cluster 4 contains a family whose secondary structure just does not fit in with any of the others.

Table 2. Structures in CRISPR Families

Family	Dot-bracket notation secondary structure (top) Consensus primary structure (bottom)
Cluster 1:	
DR-66	.....<<<<<<<...>>>>>>>............... RKUUUAAUCCCUUGUARGGAUUCUUUAKUUAUGGAAC
DR-17	......<<<<<<....>>>>>>... CUUUCUAAUCCCUYUUGGGAUUUWC
DR-25	.....<<<<<<....>>>>>>.... CUUUCaaucCCUUaUGGgauuCaUC
DR-4	.....<<<<<....>>>>>......... GUUcACuGCCGuAcAGGCaGcuuAgAAA
DR-5	...............<<<<<<<...>>>>>>>..... gUUaaAAuuaaaaAAaauCCCuAUuaGGGauuGAAAc
DR-63	..............<<<<<<<<...>>>>>>>>.... GUCAAAACACAAAAUAAUUCCCUUUGGGAAUUGAAMC
Cluster 2:	
DR-64	.............<<<<..........>>>>...... AUACGAAACGUUGAUCCAUCAAAACAAGGAUUGAGRC
DR-7	............<<<<<..........>>>>>..... guUugaGAgaAaaAuCCAcUAAAACAAGGaUuGAAAC
Cluster 3:	
DR-14	.... AUUUACAUAcCAcAUAGUUAAUAUAAAAC
DR-43	.... CUUUAUAUCCCACUACGUUCAGAUAAAAC
Cluster 4:	
DR-2	.<<<.....<<<...>>>........>>>. GuUuCAAUuCCucAaaGGuAggaUaaaAaC

This analysis points out why relying on Rfam clans to cluster families into combinable groups may not be a very good idea. Since the combined models are being used as pre-filters, there does not need to be any biological meaning to the clustering of families.

5 Conclusion

A new metric for the distance between ncRNA families based on alignment of consensus structures in a mixed primary-secondary structure alphabet has been presented. Comparison with the existing base pair conflict metric shows that this metric is potentially more effective when used for clustering of ncRNA families prior to model combination. A new model combination method was also presented which potentially finds better combined models due to use of all the available sequence data and to a single round combination instead of sequential pair-wise combination.

More analysis is needed to determine if the new model combination method is more effective since the potential strengths of the method lie in cases where more than two families are combined. This is not likely to become apparent without large-scale processing of the Rfam models into combined models.

Acknowledgments. This work was supported in part by the U.S. National Institutes of Health under Grant R15GM087646 and Grant P20RR016454.

References

1. Liu, T., Schmidt, B.: Parallel RNA Secondary Structure Prediction Using Context-free Grammars. Concurrency and Computation: Practice and Experience 17, 1669–1685 (2005)
2. Weinberg, Z., Ruzzo, W.: Faster Genome Annotation of Non-coding RNA Families Without Loss of Accuracy. In: Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology, pp. 243–251 (2004)
3. Weinberg, Z., Ruzzo, W.: Exploiting Conserved Structure for Faster Annotation of Non-coding RNAs Without Loss of Accuracy. Bioinformatics 20, 1334–1341 (2004)
4. Weinberg, Z., Ruzzo, W.: Sequence-based Heuristics for Faster Annotation of Non-coding RNA Families. Bioinformatics 22, 35–39 (2006)
5. Nawrocki, E., Eddy, S.: Query-dependent Banding (QDB) for Faster RNA Similarity Searches. PLoS Computational Biology 3, e56 (2007)
6. Smith, J.: RNA Search with Decision Trees and Partial Covariance Models. IEEE Transactions on Computational Biology and Bioinformatics 6, 517–527 (2009)
7. Smith, J.: Computational Intelligence Method to Find Generic Non-coding RNA Search Models. In: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 198–202 (2010)
8. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis. Cambridge University Press (1998)
9. Eddy, S., Durbin, R.: RNA Sequence Analysis Using Covariance Models. Nucleic Acids Research 22, 2079–2088 (1995)
10. Eddy, S.: Hidden Markov Models. Current Opinion Structural Biology 6, 361–365 (1996)

11. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic Local Alignment Search Tool. Journal of Molecular Biology 205(3), 403–410 (1990)
12. Altshcul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Research 25, 3389–3402 (1997)
13. Gardner, P., Daub, J., Tate, J., Nawrocki, E., Kolbe, D., Lindgreen, S., Wilkinson, A., Finn, R., Griffiths-Jones, S., Eddy, S., Bateman, A.: Rfam: Updates to the RNA Families Database. Nucleic Acids Research 37, D136–D140 (2009)
14. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S., Bateman, A.: Rfam: Annotating Non-coding RNAs in Complete Genomes. Nucleic Acids Research 33, D121–D124 (2005)
15. Rfam: RNA Families Database of Alignments and Covariance Models, version 9.1 (December 2008), `http://rfam.janelia.org`
16. Eddy, S.: Infernal user's guide, version 1.0.2 (2009), `http://infernal.janelia.org`
17. Nawrocki, E., Kolbe, D., Eddy, S.: Infernal 1.0: Inference of RNA alignments. Bioinformatics 25, 1335–1337 (2009)
18. Rivas, E., Eddy, S.: Secondary Structure Alone is Generally Not Statistically Significant for the Detection of Noncoding RNAs. Bioinformatics 6, 583–605 (2000)
19. Jiang, W., Wiese, K.: Combined Covariance Model for Non-Coding RNA Gene Finding. In: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2011), doi:10.1109/CIBCB.2011.5948474
20. Xu, R., Wunsch, D.: Clustering. IEEE Press Series on Computational Intelligence. Wiley (2009)
21. Lee, D., Lee, J.: Support Vector Clustering Toolbox, ver. 1.0, `http://sites.google.com/site/daewonlee/research/svctoolbox`

A Study of Compression–Based Methods for the Analysis of Barcode Sequences

Massimo La Rosa, Antonino Fiannaca Riccardo Rizzo, and Alfonso Urso

ICAR-CNR, National Research Council of Italy,
viale delle Scienze Ed.11, 90128 Palermo, Italy
{larosa,fiannaca,ricrizzo,urso}@pa.icar.cnr.it

Abstract. In this paper it is introduced a new methodology for the analysis of barcode sequences. Barcode DNA is a very short nucleotide sequence, corresponding for the animal kingdom to the mitochondrial gene cytochrome c oxidase subunit 1, that acts as a unique element for identification and taxonomic purposes. Traditional barcode analysis uses well consolidated bioinformatics techniques such as sequence alignment, computation of evolutionary distances and phylogenetic trees. The proposed alignment-free approach consists in the use of two different compression-based approximations of Universal Similarity Metric in order to compute dissimilarity matrices among barcode sequences of 20 datasets belonging to different species. From these matrices phylogenetic trees are computed and compared, in terms of topology and branch length, with trees built from evolutionary distance. The results show high similarity values between compression-based and evolutionary-based trees allowing us to consider the former methodology worth to be employed for the study of barcode sequences

Keywords: Barcode DNA, Compression–Based distances, Universal Similarity Metric, Phylogenetic trees.

1 Introduction

DNA barcoding aims at discovering and isolating a very short part of DNA of living organism for identification and taxonomic purposes [1, 2]. The very basic idea is to find and define, for each kingdom of life, such as animals, plants, fungi and so on, a single gene that works as a true "barcode" providing unique identification. In the animal kingdom, *mitochondrial gene cytochrome c oxidase subunit 1* (COI) has proven to be the best barcode sequence [3]. DNA barcoding has been used for the study of the biodiversity of several species, such as fishes, birds and some bugs [4–7].

The analysis of barcode sequences, both for identification and taxonomic purposes, is carried out by means of classic bioinformatics methodologies, based on sequence alignment and computation of dissimilarity matrices that can be used to build phylogenetic trees or to make identification of unknown species through well known threshold values [8].

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 105–116, 2013.

In this paper, an alignment-free methodological approach for the analysis of barcode sequences is proposed. It is based on compression-based distances derived from Universal Similarity Metric (USM) [9]. USM is a class of distance measures, founded on rigorous information theory concepts defined in the Kolmogorov complexity [10]. Unfortunately, Kolmogorov complexity is not computable, therefore there exists a set of USM approximations based on data compression. Compression-based methods have the advantage that they do not require a prior alignment of genomic sequences and above all they hold on strong theoretical assumptions. Evolutionary distances, in turn, are based on stochastic estimates and they do not define a distance metric.

In order to justify the use of compression–based distances for the study of barcode sequences, several datasets, belonging to different kinds of species, have been downloaded from Bold database [8]; for each dataset, a set of phylogenetic trees have been build according to the most common bioinformatics algorithms (see Section 3) a set of phylogenetic trees. Those trees, then, have been compared with phylogenetic trees obtained through state-of-the-art methods based on evolutionary distances [11].

2 Background

USM distance, as defined in [9], is "universal" in the sense it can be applied to different type of input data. In fact, it has been used for classification and clustering activities in several application domains, from text processing to language analysis, from music to image files [12]. A first attempt to use one of the USM approximations, called Normalized Compression Distance (NCD), for the study of genomic sequences has been done in [12]. The result of that work was a phylogenetic tree obtained considering complete mammalian mtDNA sequences of 24 species belonging to Eutherian order. In [13] another approximation of USM, based on GenCompress compressor [14], has been applied in order to compute a phylogenetic tree of a larger dataset containing mammalian mtDNA sequences of 34 taxa. The authors stated that USM is able to provide meaningful results when applied to very large genomic sequences and a small number of taxa.

A very important experimental assessment regarding the use of USM for different type of biological datasets has been carried out in [15]. By considering six different datasets, both of protein and genomic (complete mitochondrial genome) sequences, the authors tested two USM approximations, namely NCD and Universal Compression Dissimilarity (UCD), with several compressors in order to obtain phylogenetic trees. Those trees were then compared with gold standard taxonomies using classic tree comparison algorithms, F-measure [16] and Robinson metric [17], and they also concluded compression-based methods are allowed to be considered when dealing with biological datasets.

A different use of USM for clustering, through Self-Organizing Maps, and generation of topographic representations of bacteria datasets, considering 16S rRNA gene, was done in [18, 19], where topographic maps of three bacteria phyla were built from both evolutionary distance and NCD, showing similarities and differences between maps.

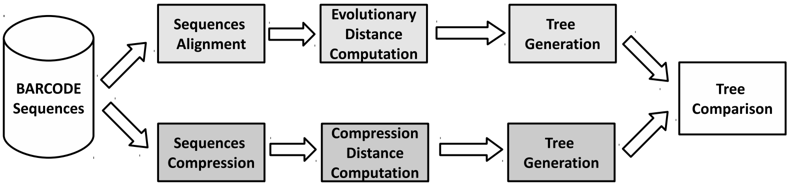


Fig. 1. Overall framework of the proposed methodology (lower workflow) compared with classic pipeline (upper workflow)

In our work, we want to demonstrate that USM, and in general compression–based distances, are also suited for the analysis of short barcode sequences, about 650 bp long, and for several datasets composed of very different species. Moreover, in order to compare phylogenetic trees obtained through evolutionary distances and compression-based methods, we adopt more recent and complete comparison tree algorithms that take into account relevant topological features of phylogenetic trees and not only basic different pairings.

3 Methods

In this Section it is presented the overall framework of our methodology; then in the following subsection, the tools and algorithms adopted in order to perform our experimental tests will be described in detail.

In Fig. 1 there are both the workflow of our proposed methodology, the lower one, and the classic workflow, the upper one, usually adopted for the analysis of gene sequences for phylogenetic purposes. After downloading barcode sequences from BOLD database [8], our approach consists in compressing the genomic sequences using GenCompress compressor [14], computing two different approximations of USM, as explained in Section 3.1, and finally building phylogenetic trees using state-of-the-art algorithms. On the other hand, classic methodology comprises sequence alignment, computation of evolutionary distance and finally generation of phylogenetic trees. Trees obtained with our and classic approach were then analyzed by means of two different tree comparison algorithms that consider different tree properties: topology and branch length.

3.1 Compression-Based Dissimilarity Measures

Universal Similarity Metric (USM) is a class of distance measures based on Kolmogorov complexity [10] and introduced by Li et al. [9]. USM allows to compare two generic data files and it has been demonstrated that it is a similarity metric, i.e the *identity axiom*, the *triangle inequality* and the *symmetry axiom* hold. The key idea of USM is to find a shared information content between two objects. Since it has been demonstrated Kolmogorov complexity is not computable, it needs to be approximated.

In our work, two different USM's approximations used for the comparison of genomic sequences have been considered: Normalized Compression Distance (NCD) [12] and the distance defined in [13] that for ease of explanation we call Information-Based Distance (IBD). In both kinds of distance, Kolmogorov complexity is approximated by means of the size of the compressed version of the sequence itself. NCD and IBD are defined respectively in Eq. 1 and Eq. 2:

$$\mathrm{NCD}(x,y)=\frac{C(xy)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}} \tag{1}$$

$$\mathrm{IBD}(x,y)=1-\frac{C(x)-C(x|y)}{C(xy)} \tag{2}$$

There $C(x)$ and $C(y)$ are the sizes, in bytes, of the compressed sequences x and y; whereas $C(xy)$ is the size of the compressed sequence obtained through the concatenation between x and y. $C(x|y)$ is the size of the compression of sequence x with respect to the reference sequence y, that is the information required to obtain x from y [20]. This kind of conditional compression is also known as vertical compression [20].

Both NCD and IBD's purpose is to find the shared information content between two sequences: NCD can be computed using a general purpose normal compressor; IBD has been introduced considering GenCompress compressor [14] to heuristically approximate Kolmogorov complexity.

GenCompress [14] is a compression algorithm optimized to work with DNA sequences. It follows the approach of Lempel and Ziv dictionary based compressors [21], taking advantage of the fact that a genomic sequence has just a four characters (a, c, g, t) dictionary. GenCompress, in fact, gives the best compression ratios only when dealing with DNA sequences: if it is applied to sequences containing more than the four nucleotide characters, it acts as a generic ascii-text compressor. GenCompress algorithm also implements a conditional version, i.e. it computes the compression of sequence x given another sequence as reference.

In this paper GenCompress is used when computing both NCD and IBD so that it is possible a direct comparison among results obtained by means of both kinds of distances.

3.2 Phylogenetic Inference

There are several methods to build a phylogenetic tree from molecular data [11, 22]. In our work two of the most used methods are considered: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [23] and Neighbor Joining (NJ) [24]. Both algorithms belong to the so called distance–based methods because they need a dissimilarity matrix among input sequences before building the tree. According to the adopted evolutionary distance model, like for instance Kimura 2–parameter [25], Tajima–Nei [26], Tamura–Nei [27], there can be different distance matrices and, consequently, different phylogenetic trees.

UPGMA is the simplest phylogenetic reconstruction algorithm, it creates an ultrametric tree (dendogram) and its basic assumption is that it builds a correct tree if the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages.

NJ considers different rates of evolution among tree's branches and it is very reliable if the input dissimilarity matrix is very close to the true evolutionary distances among sequences. NJ uses a clustering algorithm that, starting from a star topology, at each iteration pairs the nearest elements, obtaining at the end a binary tree.

3.3 Comparison of Phylogenetic Trees

In phylogenetic studies, it is possible to obtain different phylogenetic trees according to the used algorithm or the considered gene or set of genes. For this reason, several algorithms for tree comparison have been developed. The most popular is the method proposed by Robinson and Foulds [17], also known as *symmetric distance*. It computes the distance between two phylogenetic trees by considering the number of transformations, or shifts, needed to reconstruct the first tree from the second one, or vice-versa. Symmetric distance can be seen as a generalization of edit metrics [28] to phylogenetic trees.

In order to compare phylogenetic trees obtained through evolutionary distances and compression based distances, two more recent comparison algorithms, whose approach is rather different from Robinson method, have been considered: the tool presented by Nye et al. [29] and the K tree score, introduced in [30].

Nye's algorithm aims at matching branches (edges) within two trees which share similar topological features. This topological feature is the partition of leaf elements created by every branch in a tree. The similarity score for each pair of edges between two trees is given comparing the shared leaf nodes belonging to the two corresponding partitions. This process builds a sort of alignment between the two trees to compare. While Robinson metric gives each topological difference the same penalty, in Nye's algorithm different pairings have a lesser penalty if their topological features are preserved, that is they belong to the same corresponding partitions. This way similarity between trees is not expressed by a mere number of edit operations, but by considering topological properties.

This fact is better explained looking at Fig. 2, where two phylogenetic trees obtained through evolutionary distance (on the left) and compression-based distance (on the right) are shown. Thicker branches in both trees highlight a lower similarity between the corresponding subtrees; whereas thin edges identify a perfect match among the two partitions. Using Robinson metric, on the other hand, all pairs of not corresponding leaf nodes are considered as a wrong pairing, ignoring any topological feature.

K score is an extension of branch length distance (BLD) defined in [31] and it allows to obtain a similarity score depending on the similarity between branch length of both trees. Once again it differs from symmetric distance because this one does not consider branch lengths when computing the similarity score. Those two algorithms offer two kinds of comparison: Nye tool gives a score based only on the similarity between trees topologies; K score takes into account the

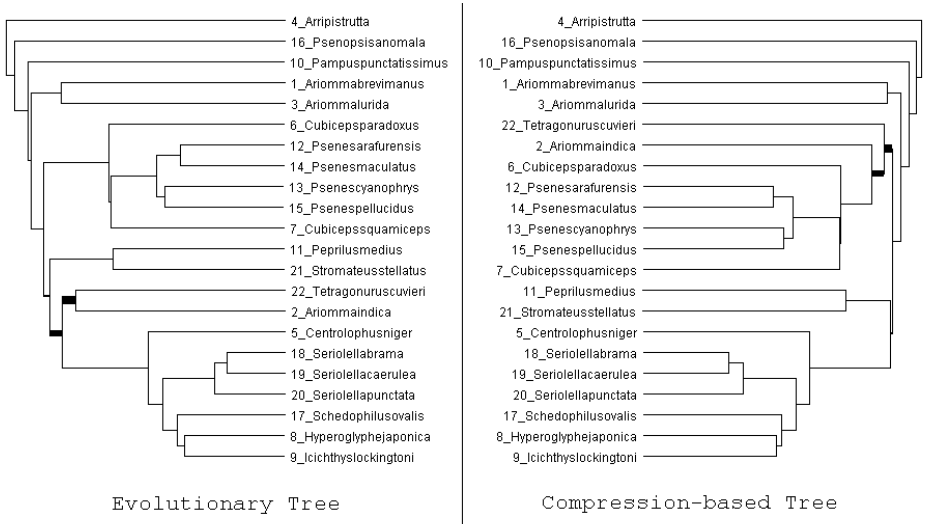


Fig. 2. Comparison between phylogenetic trees obtained from evolutionary distance (left) and compression-based distance (right). Thicker edges mean the corresponding partitions within the two tress have not exactly the same leaf nodes.

similarity between branch length's trees. This way we can test our results both in terms of topology and branch length similarity.

4 Results

In this section we report experimental tests used for evaluating two compression-based algorithms (NCD and IBD). The evaluation is based on the comparison of phylogenetic trees generated with both UPGMA and NJ algorithms.

4.1 Dataset Description

In order to test the performance of the discussed compression-based algorithms, we used 20 datasets from "Barcode of Life Data" Systems (BOLD) Project [8]. Among more than 1000 available datasets, we considered a subset composed by those datasets that respect two main criteria: first, for each dataset all the sequences (representing species or specimens) are the mitochondrial COI-5P gene and second, all the datasets belong to different *familia* of the animal kingdom. From this subset, we randomly selected 20 datasets.

All datasets used during experimental tests are reported in Table 1. The first column shows datasets acronyms, as reported in BOLD database. For each dataset, Table 1 reports four features: the number of specimens and species (respectively second and third column); the number of sequences having at least

Table 1. 20 Datasets selected from Barcode of Life Data System. Some datasets are clustered in 5 groups of distinctive features.

	Dataset	#Specimens	#Species	Sequences with	Length of	Distinctive features				
		(sequences)		undefined bases	sequences	G1	G2	G3	G4	G5
1	**JTB**	225	53	1/225	658-899	✓				
2	**DLTC**	67	40	1/67	689-1821	✓				
3	**Onychophora**	210	52	2/210	451-884	✓				
4	**AGWEB**	33	33	29/33	460-890	✓	✓			✓
5	**GBFCJ**	202	61	14/202	537-1446	✓	✓			
6	**GZPSE**	78	23	6/78	601-658		✓			
7	**RDMYS**	37	6	12/37	636		✓	✓		
8	**ARCPU**	52	28	3/52	901			✓		
9	**BRBP**	106	17	0/106	658			✓	✓	
10	**AGFDO**	22	22	0/22	901			✓	✓	✓
11	**AECI**	30	30	0/30	605-679				✓	✓
12	**SIBHI**	85	38	0/85	673-694				✓	
13	**BLSPA**	86	86	4/86	604-658					✓
14	**ABSMC**	72	46	1/72	650-657					
15	**AGFSU**	48	42	1/48	605-680					
16	**AGLUO**	46	38	1/46	633-639					
17	**DSALA**	44	12	5/44	649-651					
18	**FBLOT**	64	34	2/64	419-658					
19	**MJMSL**	198	76	9/198	559-658					
20	**WXYZ**	34	9	1/34	650-680					

one undefined base (forth column) and the range of sequences' length (fifth column).

The last column is composed by 5 sub-columns that indicate a particular set of features. The meaning of each group is reported in the following, whereas the analogies among these datasets will be investigated in the next Section:

– G1: Datasets in this group contain some sequences much longer than the other ones of the same dataset;
– G2: In these datasets there is an high percentage of sequences with undefined bases;
– G3: All the sequences in these datasets have the same length;
– G4: Sequences in these datasets do not have undefined bases;
– G5: These datasets contain sequences with one specimen for each species.

The BOLD system provides, for each dataset, a distance matrix obtained by default using the Kimura 2-parameter distance model. With regards to the compression-based algorithms, since they require as input a list of sequences in order to generate the distance matrix, we also downloaded a list of (pre-aligned) COI-5P gene sequences for each dataset.

Table 2. Similarity and K-score among phylogenetic trees: evolutionary technique with Kimura 2-parameter distance versus compression based algorithms (both NCD and IBD)

	Dataset	**Tree similarity (Nye et al.)**				**K-Score**			
		NCD		**IBD**		**NCD**		**IBD**	
		UPGMA	**NJ**	**UPGMA**	**NJ**	**UPGMA**	**NJ**	**UPGMA**	**NJ**
1	**JTB**	0.75	0.61	**0.85**	0.59	0.1852	—	0.2090	—
2	**DLTC**	**0.86**	0.79	0.84	0.77	0.6842	—	0.6973	—
3	**Onychophora**	0.89	0.77	**0.92**	0.81	0.1165	—	0.1333	—
4	**AGWEB**	0.73	0.76	0.77	**0.89**	0.0667	0.0674	0.0775	0.0779
5	**GBFCJ**	0.80	0.72	**0.82**	0.72	0.1170	—	0.1155	—
6	**GZPSE**	0.84	**0.86**	0.85	0.85	0.0588	—	0.0714	—
7	**RDMYS**	0.78	0.60	0.81	**0.87**	0.0472	—	0.0587	—
8	**ARCPU**	0.87	**0.94**	0.87	0.87	0.0562	—	0.0720	—
9	**BRBP**	**0.99**	0.89	**0.99**	0.82	0.0772	—	0.1149	—
10	**AGFDO**	**0.92**	0.88	**0.92**	0.88	0.0315	0.0323	0.0607	0.0632
11	**AECI**	0.89	0.82	**0.90**	0.82	0.0517	—	0.0839	—
12	**SIBHI**	0.92	0.90	**0.94**	0.90	0.0581	—	0.0976	—
13	**BLSPA**	**0.88**	0.82	0.85	0.79	0.0424	0.0485	0.0621	0.0672
14	**ABSMC**	**0.97**	0.93	0.92	0.95	0.0720	—	0.1211	—
15	**AGFSU**	0.84	0.84	**0.89**	0.85	0.0609	0.0598	0.0910	0.0954
16	**AGLUO**	0.97	0.90	**0.98**	0.90	0.0394	0.0442	0.0615	0.0646
17	**DSALA**	**0.91**	0.88	**0.91**	0.88	0.0706	—	0.0940	—
18	**FBLOT**	0.89	0.81	**0.90**	0.82	0.0655	—	0.1011	—
19	**MJMSL**	**0.88**	0.82	**0.88**	0.79	0.0912	—	0.1211	—
20	**WXYZ**	0.92	0.79	**0.95**	0.76	0.0629	—	0.0926	—

4.2 Experimental Tests

Experimental tests aim to evaluate the quality of phylogenetic reconstructions obtained by means of compression-based algorithm. Using aforementioned evaluation techniques, we compare phylogenetic trees that have been generated with the two most used algorithms: UPGMA and NJ.

Results are reported in Table 2. This table is composed by three main columns: the first one contains dataset acronyms, the second one the "Tree Similarity" scores and the last one the "K Score". Second and third columns, in turn, contain sub-columns in order to show results obtained with UPGMA and NJ trees with both NCD and IBD distances. In the "Tree Similarity" columns, each number represents a percentage value, where "1" means trees have the same topology, that implies the compression-based algorithm preserves the evolutionary taxonomy. Numbers in bold type are the best scores for each dataset. The "K Score" column, instead, reports values that measure the difference between two trees in terms of branch length. In this case, lower values mean higher similarity among trees. Unfortunately, NJ algorithm sometimes generates trees with negative branches that can not be computed by K score algorithm: in fact, although

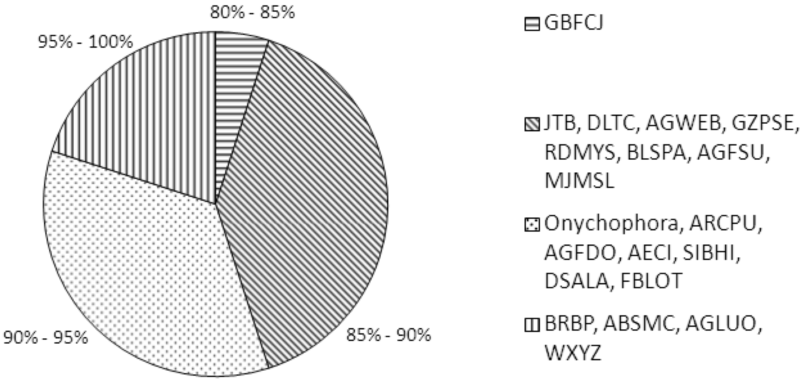


Fig. 3. Pie chart representing the percentage of similarity between evolutionary and compression-based tree for each datasets

mathematically the neighbor joining algorithm is admitted to produce negative values, biologically a tree with some negative branches is meaningless [31]. In this situation, a "–" symbol is reported.

First of all, considering the tree similarities results, we can state that the compared trees are quite similar. More in detail, considering the best values for each dataset, we obtained the pie chart in Fig. 3. The most of datasets are between 85% and 95% and four of them have a topological similarity greater than 95%. Only the fifth dataset (GBFCJ) gives the a result of 82%.

These results are not so surprising because the quality of retrieved datasets are different each other, looking at the five groups of datasets reported in Table 1. Group G1 contains some sequences much longer than the other ones: for these datasets, compression-based similarity algorithms give poor results, since they take into account mutual length of sequence.

Group G2 is composed by four datasets with several sequences that contain some special symbol (i.e. Y or N) to represent undefined nucleotides. In this case, as previously said in Section 3, the GenCompress algorithm works as a generic compressor of ASCII string, reducing its performance. For instance, the dataset GBFCJ, belonging both to first and to second group, shows the lowest value of tree similarity (82%). Another dataset belonging to G1 and G2 groups, is AGWEB. This dataset, with respect to GBFCJ, has an higher value of tree similarity (89%), because it has a lower spread in sequences length.

Groups G3 and G4 in Table 1 contain those datasets with respectively the same length for all sequences and with a complete COI-5P gene sequencing. BRBP dataset belongs to both these groups and represents the best one among datasets used in this paper, since it has no sequences with undefined bases and all the sequences have the same length of 658-bp, representing COI-5P gene length proposed as a potential 'barcode' in [2]. This dataset, composed of 106 elements, reaches a value of 99% tree similarity with its corresponding evolutionary tree. It is interesting to notice that the other datasets having sequence length close to

658-bp, and that do not belong to the first or the second group, score the best results, such as AGLUO (98%) and SIBHI (94%).

Group G5 reports datasets with a single specimen (sample) for each species. In terms of tree similarity, datasets with only a specimen for each species do not produce better results than datasets with more than a specimen for each specie, with both compression-based distances. In other words, compression algorithm works fine also at specimens level.

Considering the type of compression-based distances, obtained results demonstrate in the most of case (75%) IBD reaches highest values in terms of tree similarity, especially when UPGMA is used for generate trees, with the exception of datasets in G1 of Table 1. In fact, for datasets with an high percentage of sequences with undefined bases, NJ is able to better represent the evolutionary tree, for instance AGWEB has 87% of undefined bases and reaches the better value of similarity (89%) with IBD and NJ algorithm. This means that in all cases IBD algorithm is able to preserve the topology of an evolutionary tree of DNA barcode sequences.

As for the K-score column in Table 2, results confirm all the considerations previously said, except for compression distance algorithm analysis. In fact, in terms of differences in the relative tree branch length, it appears NCD algorithm works better than IBD. It is possible to notice that datasets in group G1 of Table 1 score lesser results, e.g. DLTC (0.684), whereas datasets in group G4 score the best results, e.g. AGFDO (0.031).

5 Conclusion

In this paper we presented a deep analysis about the use of compression-based methods, such as NCD and IBD, for the study of short DNA barcode sequences. NCD and IBD are both approximations of Universal Similarity Metric, that is a class of general-purpose distances based on non-computable Kolmogorov complexity. In previous works, USM and its approximations have been applied in the case of the analysis of complete mitochondrial genome of few species: there they showed how phylogenetic trees obtained through USM had a very similar topology to those ones obtained through classic bioinformatics methods based on sequence alignment and evolutionary distances computing. By employing compression-based methods there is no need to align input sequences and moreover USM represents a distance metric, whereas evolutionary distances are stochastic distance estimates lacking metric properties such as triangle inequality. In this work we extended the use of NCD and IBD to DNA barcode sequences, typically 650 bp long. We compared phylogenetic trees of 20 datasets obtained from NCD and IBD, using NJ and UPGMA algorithms, with trees of the same datasets obtained from Kimura 2-parameter evolutionary distance. The comparison was done by means of two different algorithms, considering both topological and branch length similarities. The results we presented show that trees obtained from compression-based methods are very similar (above 90%), and in some cases equal, to the ones built from classic distance. In few situations, characterized by some flaws in input datasets, we obtained similarity scores of about 85%, demonstrating compression-based methods are robust enough to deal with noisy datasets. In the near future we are going to provide other comparisons between trees using other kinds of evolutionary distances and phylogenetic reconstruction algorithms so that we can definitively use compression-based methods for the study of phylogenetic relationships with DNA barcode sequences.

References

1. Savolainen, V., Cowan, R.S., Vogler, A.P., Roderick, G.K., Lane, R.: Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1805–1811 (2005)
2. Hebert, P.D.N., Cywinska, A., Ball, S.L., de Waard, J.R.: Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321 (2003)
3. Hebert, P.D.N., Ratnasingham, S., de Waard, J.R.: Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. Biol. Sci. 270(suppl. 1), 96–99 (2003)
4. Costa, F.O., Carvahlo, G.R.: The Barcode of Life Initiative: synopsis and prospective societal impacts of DNA barcoding of fish. Genomics, Society and Policy 3, 29–40 (2007)
5. Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., Francis, C.M.: Identification of Birds through DNA Barcodes. PLoS Biol. 2(10), e312 (2004)
6. Smith, M.A., Fisher, B.L., Hebert, P.D.N.: DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Phil. Trans. R. Soc. B 360, 1825–1834 (2005)
7. Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W., Hebert, P.D.N.: DNA barcodes distinguish species of tropical Lepidoptera. PNAS 103(4), 968–971 (2006)
8. Ratnasingham, S., Hebert, P.D.N.: BOLD: The Barcode of Life Data System. Molecular Ecology Notes 7, 355–364 (2007)
9. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.M.B.: The Similarity Metric. IEEE T. Inform. Theory 50(12), 3250–3264 (2004)
10. Li, M., Vitanyi, P.M.B.: An Introduction to Kolmogorov Complexity and its Applications, 2nd edn. Springer, New York (1997)
11. Makarenkov, V., Kevorkov, D., Legendre, P.: Phylogenetic network construction approaches. Applied Mycology and Biotechnology 6, 61–97 (2006)
12. Cilibrasi, R., Vitanyi, P.M.B.: Clustering by Compression. IEEE T. Inform. Theory 51(4), 1523–1545 (2005)
13. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17(2), 149–154 (2001)
14. Chen, X., Kwong, S., Li, M.: A compression algorithm for DNA sequences. IEEE Engineering in Medicine and Biology Magazine 20(4), 61–66 (2001)
15. Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., Valiente, G.: Compression-based classification of biological sequences and structures via the Universal Similarity Metric: Experimental assessment. BMC Bioinformatics 8(252) (2007)
16. van Rijsbergen, C.J.: Information Retireval. Butterworths, London (1979)
17. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical Biosciences 53(1), 131–147 (1981)

18. La Rosa, M., Rizzo, R., Urso, A., Gaglio, S.: Comparison of Genomic Sequences Clustering Using Normalized Compression Distance and Evolutionary Distance. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 740–746. Springer, Heidelberg (2008)
19. La Rosa, M., Gaglio, S., Rizzo, R., Urso, A.: Normalised compression distance and evolutionary distance of genomic sequences: comparison of clustering results. Int. J. Knowledge Engineering and Soft Data Paradigms 1(4), 345–362 (2009)
20. Grumbach, S., Tahi, F.: A new challenge for compression algorithms: genetic sequences. J. Information Processing and Management 30(6), 866–875 (1994)
21. Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Trans. Inform. Theory 23(3), 337–343 (1977)
22. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, New York (2000)
23. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy: The Principles and Practice of Numerical Classification. W.H. Freeman, San Francisco (1973)
24. Saitou, N., Nei, M.: The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. 4(4), 406–425 (1987)
25. Kimura, M.: Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. 78, 454–458 (1981)
26. Tajima, F., Nei, M.: Estimation of evolutionary distance between nucleotide sequences. Molecular Biology and Evolution 1, 269–285 (1984)
27. Tamura, K., Nei, M.: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10, 512–526 (1993)
28. Atallah, M.J., Blanton, M.: Algorithms and Theory of Computation Handbook. CRC Press LLC (1999)
29. Nye, T.M.W., Liò, P., Gilks, W.R.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22(1), 117–119 (2006)
30. Soria-Carrasco, V., Talavera, G., Igea, J., Castresana, J.: The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23(21), 2954–2956 (2007)
31. Kuhner, M.K., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468 (1994)

Assessing Agreement between microRNA Microarray Platforms via Linear Measurement Error Models

Niccolò Bassani, Federico Ambrogi, and Elia Biganzoli

University of Milan
{niccolo.bassani,federico.ambrogi,elia.biganzoli}@unimi.it
http://www.unimi.it

Abstract. Over the last years miRNA microarray platforms have provided insights in the biological mechanisms underlying onset and development of several diseases and have thus become a very popular instrument for profiling thousands of miRNA simultaneously. However, because of large variety of microarray platforms available, an assessment of their performance in terms of both within-platform reliability and between-platform agreement is needful. In particular, assessment of platform concordance has been a very relevant issue in the past decade. To date, only a few studies have evaluated this problem in the field of miRNA microarray, and mostly by using improper statistical methods such as the Pearson and Spearman correlation coefficients. In this work we suggest to use a recently proposed modified version of the classical Bland-Altman approach for comparing clinical measurement methods. This modified version is useful in that allows not only to evaluate agreement between different miRNA microarray platforms, but also to assess which are the potential sources of disagreement/bias between them.

Two samples were profiled using Affymetrix, Agilent and Illumina miRNA platform using three technical replicates each, and pairwise agreement between platforms was evaluated within each sample. Our results suggest that, after bias correction, Illumina and Agilent show the best patterns of agreement for both samples involved in the experiment, whereas Affymetrix is the one which seem to "disagree" most, suggesting that a linear relationship as that hypothesized by the measurement error model used is not able to capture the complexity of the phenomenon.

In the future it will be interesting to apply this method also to the comparison of microarray and NGS platform, a topic which is becoming more and more relevant, also by adopting non-linear measurement error models to depict relationships between platforms.

Keywords: microRNA,microarrays, agreement, measurement error model, Bland-Altman.

1 Introduction

MicroRNAs are small non-coding RNA molecules which have been shown to play a critical role in tumorigenesis [1–4] and in several other pathologies [5–7].

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 117–131, 2013.

In order to measure miRNA intensity levels, several methods, such as RT-qPCR, high-throughput sequencing and microarrays, have been developed that enable researchers to profile thousands of miRNAs simultaneously across different experimental conditions [8]. MiRNA microarrays in particular, since their first appearance in 2004 [9], have known a relevant spread in life sciences and are now routinely used in bio-molecular laboratories.

Nonetheless, there's been a substantial lack of evaluation of between-platform agreement This issue has a long history which dates back to almost ten years ago, when researchers first tried to evaluate comparability between different cDNA platforms, obtaining quite disheartening results which were mainly caused by lack of standardized protocols for management and processing of this ”new” kind of data. Subsequent studies showed that patterns of reproducibility between platforms were much stronger than what had initially been found, and microarray platforms became a standard tool in most laboratories [10].

In the context of miRNA microarrays, only few studies have attempted at evaluating between-platform concordance [11, 12]. Results were reported mainly as correlation coefficients (both Pearson and Spearman) for evaluating inter-platform performance, calculated on a subset of microRNA which depended on detection calls concordant between platforms, i.e. miRNA which were called as ”detected/present” on all platforms considered for the analysis. Additionally, Sato and colleagues assessed between-platforms reproducibility also in terms of miRNAs which were commonly differentially expressed between samples for all platforms [11].

As a matter of fact, many papers dealing with this issue in gene expression suffered from several limitations which have propagated also in miRNA studies, both in the pre-processing of data and in the statistical methods adopted. For the pre-processing, for instance, most work filter data according to some intensity-based criterion, thus excluding from the analysis genes/miRNA which are ”switched-off”, posing a relevant problem: which is their expression profile between platforms in samples where they are ”turned on”? Excluding genes because of their low intensity can severely bias results in both directions, by over- and under-estimating real patterns of agreement/reproducibility. In terms of statistical methods, instead, the most reported statistical ”measure” to assess concordance (or comparability, as it is called in some papers) is the correlation coefficient (Pearson and/or Spearman), a measure that has been already critically discussed with respect to its appropriateness to evaluate reproducibility [13]. That is, an index which aims at quantifying the linear relationship between two variables X and Y is used to tell whether X and Y show some pattern of concordance.

Moreover, all of the works published on this topic considered the intensity values as unaffected by some measurement error, that is assumed that no difference existed between ”real” and ”observed” value. Such an assumption has a direct implication also on the computation of simple correlation coefficient, since the relationship between X and Y tends to become milder if some measurement error is present and is not considered [14]. In addition to this, we believe that evaluating platform agreement is a complex issue that can not be reduced simply to the computation of some index, mainly because it is not only important to assess agreement/concordance between arrays, but also, and probably even more, eventual disagreement, its sources and how to correct for it. For this reason, we suggest to use a modified version of the classical Bland-Altman approach for assessing agreement [15, 16], a well-known approach in the context of clinical and laboratory research. This modified version [18] takes into account measurement error to evaluate sources of possible disagreement between two measurement methods (in our case microarray platforms) and to correct for it according to a linear measurement error model . The paper is organized as follows: in section 2 we present platforms and sample features, as well as experimental design and pre-processing of data. Moreover, we illustrate the modified Bland-Altman methods; in section 3 main results of the study are briefly presented; in section 4 advantages and drawbacks over other approaches are critically discussed and in section 5 we summarize main results of the study and outline possible future developments of this work.

2 Materials and Methods

2.1 miRNA Microarray Platforms

Affymetrix GeneChip© miRNA Array. The Affymetrix GeneChip© miRNA Array (Affymetrix Santa Clara, CA, USA) includes 46227 perfect match (PM) probes, representing 7815 probesets, of which only 847 (10.83%) represents human microRNA; the remaining act as control probe sets which are expected not to be expressed when human samples are hybridized. This annotation refers however to an old version of the array, since, at present time, more than 1000 microRNAs have been discovered and characterized [17]. The array contains 4 identical probes for each miRNA, located in specific spots on the array, with a length between 16 and 25-mer.

Intensity .CEL files were obtained from the scan images and imported to Affymetrix© miRNA QC Tool software (Version 1.0.33.0) to quantify the signal value. We assessed QC by plotting the average intensity of the oligo spike-in and background probe sets (included in Control target content) across all the arrays. According to Genisphere, oligo spike-in 2, 23, 29, 31 and 36 probe sets should present a value of more than 1000 intensity units to accept array quality. The miRNA arrays were detected using Affymetrix detection algorithm, based on non parametric Wilcoxon Rank-Sum test, applied independently on each array and probe/probe set; a p-value greater than 0.06 stands for not detected above background [19]. For data normalization, we chose default method obtaining log2 expression values (expression values data matrix) from the raw data (intensity values data matrix). Briefly, this method involves the following three steps: grouping the background probes intensities based on GC content, where the median intensity of each bin was the correction value for each probe with the same GC content; a quantile normalization and, finally, a median polish summarization. To obtain a single intensity value for each microRNA mapped on the array log2 intensity measures for replicated spots were averaged.

Agilent Human miRNA Microarray (V1). The Agilent Human miRNA Microarray (V1) platform contains information on 961 microRNAs, of which 851 (88.55%) are human microRNAs, for each of which there's a variable number of replicated probes mapping to it. Notably, these probes are not all different, and there are sub-levels of replication within each set of replicated probes: that is, out of 16 replicated probes (constituting a probeset) for miRNA A there could be four groups of 4 probes with the same nucleotide sequence, for instance. Of all the microRNAs, both human and non-human, 97.7% have 16 replicates, and this includes all the human microRNAs. One miRNA shows 182 replicates and is labelled as `NegativeControl` on the array, representing the level of background noise. In terms of probe replication within each probeset, if we consider only human microRNA we find that 308 miRNAs (36.2%) are "interrogated" by 4 different sequences, 56 (6.6%) by 3 sequences, 483 (56.8%) by 2 sequences and 4 (0.5%) by a single sequence.

Images were scanned using the Agilent Feature Extraction (AFE) software (Version), obtaining the TotalGeneSignal (TGS) for all microRNAs on the array. Negative values were transformed by adding the absolute value of the minimum TGS intensity in the experiment as extracted by the AFE + 2 before log2 transformation [20]. Data extracted from AFE were imported in the R environment [21] and processed using the AgiMicroRna package, available in Bioconductor [22, 23].

Illumina humanMI_V2. Illumina humanMI_V2 platform contains 1145 miRNAs of which 858 (74.93%) are human miRNAs. Each miRNA is quantified by a large series of copies of the same oligonucleotides which are attached to a "bead". This bead-level information is then collapsed into a single intensity measure for each microRNA by means of a proprietary algorithm by Illumina, Inc. Raw data were processed using the proprietary BeadStudio software (Version 3.3.8), and background subtraction was performed according to the method developed by Irizarry et al. for Affymetrix microarrays [24].

2.2 Samples

The two samples involved in the study are a renal tumor cell line named A498 (ATCC, Manassas, VA, USA) [25] and a pool of twenty different human normal tissues (namely hREF), obtained from the First Choice© Human Total RNA Survey Panel, (Ambion Inc, Austin, TX, USA). RNA material was analyzed in different labs, as follows: Affymetrix processing took place at the Biomedical Technologies Institute of the University of Milan (Segrate, Italy), Illumina and Agilent processing was performed at the Department of Experimental Oncology of the National Cancer Institute (Milan, Italy). For both samples three technical replicates for each platform were performed, leading to a total of 18 arrays, 6 for each different microarray platform.

2.3 miRNA Selection and Normalization

We selected human miRNAs common to all platforms according to their name and confirmed by search on miRBase (Release 18, November 2011). Unlike other published works, we did not filter miRNAs on detection basis because such an approach could possibly introduce a bias in results. That is, some of the miRNAs which are filtered out because are "switched-off" could show patterns of within- and/or between-platform disagreement in another experiment where they are "turned-on", thus leading to an over-estimate of the level of reliability. The choice of considering only human microRNAs should circumvent this issue, and at the same time provide relevant information since human microRNAs are commonly those that are of major interest in biomolecular investigation.

Moreover, no data normalization was performed. Generally, almost all works who focused on comparing microarray platforms applied some normalization to their data (for instance [11, 12]), but this is a non trivial issue which is to be carefully evaluated since, to date, normalization for miRNA microarray has been largely debated, with results that have been somehow discordant (see for instance [26–28]), so that no "gold-standard" methods exists. Additionally, normalizing data in the context of assessing platform agreement poses some other relevant problems. That is, if we normalize data on two different platforms and then compare normalized data, we are not simply assessing concordance/agreement/reproducibility between platforms, but evaluating a sort of normalization/platform interaction to understand whether that specific normalization leads to concordant data. So, we could find high level of between-platforms agreement due not to the platforms themselves but to the normalization used or, on the other hand, the same normalization on different platforms could highlight patterns of discordance that can not be ascribed to the platforms. Nonetheless, comparing un-normalized data exposes to the risk of finding poor concordance because of incidental batch effects occurred in the experiment which may lead to an under-estimate of the "real" agreement between platforms. In this paper we have chosen to use non normalized data, so that we could assess performance of different platforms "per se".

2.4 Between-Platform Agreement

Agreement between platforms was evaluated using a modified version of the Bland-Altman approach. Such a modification, suggested by Liao *et al.* [18] in a non-genomic study context, allows not only to assess whether two methods of measurement are concordant but also to provide information on the eventual sources of disagreement. In a nutshell, the method involves the estimation, for each platform pair and separately for each sample, of a measurement error model, i.e. a model where also the independent variable(s) X are assumed to be affected by some uncertainty, of the form:

$$Y_i = a_0 + b_o X_i^0 + \epsilon_i \tag{1}$$

$$X_i = X_i^0 + \delta_i \tag{2}$$

where $(X_i^0, Y_i^0), i = 1, ..., n$ are the unobserved true values of the two measurement methods to be compared, i.e. miRNAs intensities on the two platforms, a_0 and b_0 are the intercept and the slope of the model, which conceptually have the same meaning and interpretation as in the OLS linear model. The ϵ_i and δ_i are the i.i.d. error components of the model, which follow a normal distribution with 0 mean and variances σ_ϵ^2 and σ_δ^2, respectively. To estimate this model one needs to know the ratio λ of the error variances of X and Y, possibly by means of replication or, when replication is not feasible, by setting it equal to 1, thus assuming equal error variances for both methods. In this study we have evaluated both strategies, using the technical replicates to estimate λ by fitting a linear model using the factor "replicate" as independent variable and considering the estimated residual variance as the sample error variance for the platform. Once the parameters of the model are estimated, modified versions of the agreement interval for $Y - X$, which is here assumed to follow a Normal distribution with mean $a_0 + (b_0 - 1)X_i^0$ and standard deviation $\sqrt{(1+\lambda)}\sigma_\delta$, that was proposed by Bland & Altman [14, 16] are estimated according to the bias (fixed or proportional) one wishes to correct for when comparing two platforms, and results are visualized graphically. Under condition of perfect agreement (i.e. a = 0 and b = 1), then the new agreement interval has the following form:

$$\Delta = \left(-t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta, +t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta\right) \tag{3}$$

However, in some cases it is known that some bias exists between the two methods, both fixed and/or proportional. In the first case, i.e. when only a fixed bias between the two methods is present, this is equivalent to a situation when the intercept in the model is different from 0 (i.e. the confidence interval for the parameter does not contain the value 0) and the slope b_0 of the model is not different from 1 (i.e. the value 1 is included in the relative confidence interval). This means that the two methods possibly differ by a fixed shift, which is represented by the estimate of a_0, which leads to the following modified agreement interval:

$$\Delta = \left(a_0 - t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta, a_0 + t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta\right) \tag{4}$$

Similarly, the two methods could differ only by a proportional bias described by the slope of the model (i.e. $a_0 = 0$ and $b_0 \neq 1$), leading to a modified agreement interval which has the following form:

$$\Delta = \left((b_0 - 1)X_i - t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta, (b_0 - 1)X_i + t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta\right) \tag{5}$$

Finally, the two methods could differ both by a fixed and a proportional bias (i.e. $a_0 \neq 0$ and $b_0 \neq 1$), which leads to this interval:

$$\Delta = \left(a_0 + (b_0 - 1)X_i - t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta, a_0 + (b_0 - 1)X_i + t_{1-\alpha/2,n-1}\sqrt{1+\lambda}\hat{\sigma}_\delta\right) \tag{6}$$

Notably, the X_i are the actual measured values of the "reference" method X. Once the proper interval has been computed, depending on the inference on the parameters, one can visualize the results by plotting the differences between measurement methods versus the sample labels (when no bias or only fixed bias is present) or versus the sample labels ordered according to the values of the X method (when only proportional or both biases are present), then adding the lines representing the agreement interval and evaluating how many subjects lie within the interval. If no more than a predefined number of k subjects show $Y - X$ differences that lie outside these intervals than the two methods are said to be in agreement. The choice of the threshold k depends on the tolerance one wishes to accept: the lower the tolerance for disagreeing subjects, the lower the value of k.

Since our aim is to assess global agreement of expression profiles between platforms rather than evaluating single miRNAs concordance between arrays, we considered the microRNAs to be the subjects and the different platforms to be the measurement methods, and compared all array pairs separately for each cell line.

3 Results

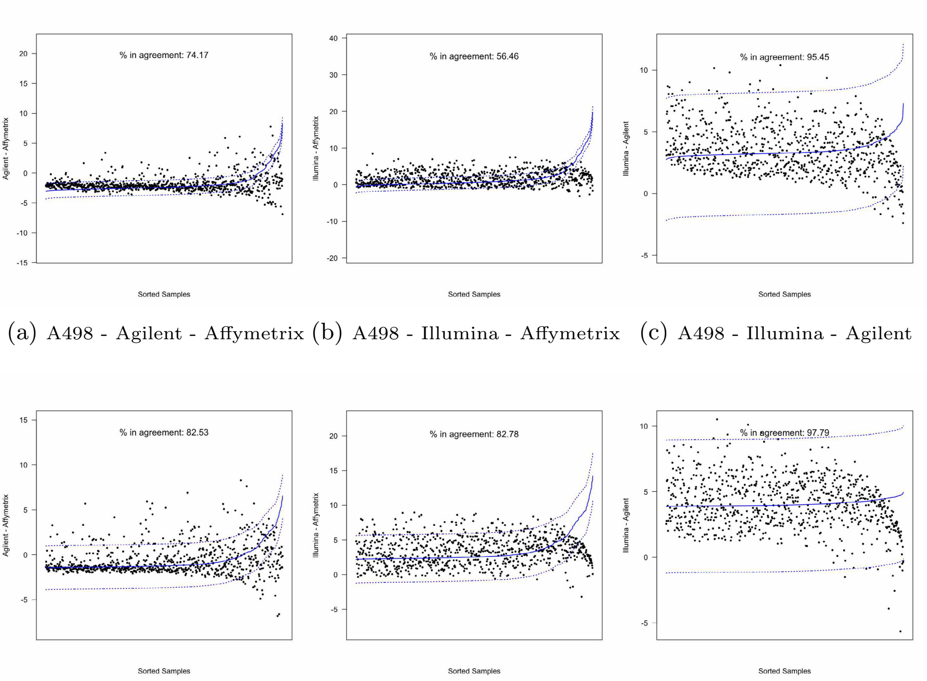
To perform agreement evaluation, we averaged miRNAs intensities across technical replicates for each array and for each sample. Then, we evaluated pair-wise arrays agreement in terms of miRNA lying within the modified agreement interval described in section 2.4. Estimates of the measurement error model for error-variance ratio λ equal to 1 are presented in table 1.

Table 1. Estimates of the linear measurement error model, $\lambda = 1$

		a_0		b_0	
	Pair	*Estimate*	CI95%	*Estimate*	CI95%
A498	Agilent vs Affymetrix	-12.4128	(-16.9575 , -11.8681)	2.8265	(2.7461 , 2.9068)
	Illumina vs Affymetrix	-17.4064	(-18.6406 , -16.1722)	4.2889	(4.1679 , 4.4098)
	Illumina vs Agilent	2.1916	(1.7773 , 2.6058)	1.3254	(1.2407 , 1.4100)
hREF	Agilent vs Affymetrix	-6.1037	(-6.4471 , -5.7603)	1.9610	(1.8964 , 2.0255)
	Illumina vsAffymetrix	-4.7630	(-5.5006 , -4.0254)	2.4418	(2.3472 , 2.5363)
	Illumina vs Agilent	3.6033	(3.3791 , 3.8274)	1.0925	(1.0371 , 1.1479)

It appears that the relationship between Agilent and Illumina is the one which is closest to the agreement line with intercept 0 and slope 1 for both samples, whereas models which include the Affymetrix platform for line A498 show a very negatively large intercept (-12,4128 and -17.4064) which possibly reflects the technical bias already highlighted for this samples. However, if we consider line hREF we can note that also in this case Affymetrix is the array which deviates most from the line of perfect agreement with both Illumina and Agilent, whereas these two show patterns very close to concordance (slope = 1.0925, CI95%: 1.0371 - 1.1479).

Since the confidence intervals suggest an intercept different from 0 and a slope different from 1 for all comparisons in both samples, we build the agreement intervals following formula 6, and visualize results for lines A498 and hREF in figure 1. At the top of each graph we have reported the proportion of microRNA which are found to be in agreement after bias correction, i.e. miRNA that lie within the agreement interval.



(a) A498 - Agilent - Affymetrix (b) A498 - Illumina - Affymetrix (c) A498 - Illumina - Agilent

(d) hREF - Agilent - Affymetrix (e) hREF - Agilent - Affymetrix (f) hREF - Illumina - Agilent

Fig. 1. Agreement intervals for the modified Bland-Altman method, $\lambda = 1$

If we consider A498 comparisons (upper panels) results are substantially worse when Affymetrix platform is considered: 74.17% (CI95%: 71.02 - 77.15) of microRNA, i.e. 603, lie within the agreement interval for the comparison Agilent - Affymetrix, 56.46% (CI95%: 52.97 - 59.90), i.e. 459, for the comparison Illumina - Affymetrix. Comparing Illumina and Agilent results in 95.45% (CI95%: 93.78 - 96.78), i.e. 776, microRNA in the agreement interval. hREF comparisons, on the other hand, show better patterns of agreement after bias correction, resulting in 82.53% (CI95%: 79.75 - 85.08) of microRNA, i.e. 671, within the agreement interval for the comparison Agilent - Affymetrix, 82.78% (CI95%: 80.01 - 85.31) for the comparison Illumina - Affymetrix, i.e. 673, and 97.79% (CI95%: 96.52 - 98.68) for the comparison of Illumina and Agilent, i.e. 795 microRNA out of 813. Confidence intervals for the proportions were computed using the Clopper-Pearson exact method [29].

Yet, these results of this modified version of the Bland-Altman plot showed so far rely on the validity of the assumption $\lambda = 1$ which, though reasonable in many practical situations, could be misleading in the context of microarrays. That is, assuming an error variance which is the same for very different platforms may be not the best choice. For this reason, we have fitted a random effects model for each platform and sample, and the estimates of λ were obtained by computing proper ratios of the residual error variances (see table 2). We can see that the estimates for the comparison Illumina - Agilent are quite close to 1 for both samples (1.125 (CI95%: 1.039 - 1.218) for hREF and to 1.352 (CI95%: 1.248 - 1.463)), whereas the estimates for the other comparison are quite different from 1, in particular for Affymetrix-related comparisons in A498 line. Given these estimates, we expect the results for the concordance analysis to be substantially similar for the Illumina - Agilent comparison, whereas relevant differences are expected for the remaining pairs, in particular for samples A498.

Table 2. Estimates of λ and CI95%. Values obtained as ratio of σ_ϵ^2 (error variance of Y) and σ_δ^2 (error variance of X).

		λ	CI95%
A498	*Agilent - Affymetrix*	4.125	3.810 - 4.466
	Illumina - Affymetrix	5.576	5.150 - 6.037
	Illumina - Agilent	1.352	1.248 - 1.463
hREF	*Agilent - Affymetrix*	2.608	2.409 - 2.824
	Illumina - Affymetrix	2.935	2.711 - 3.178
	Illumina Agilent	1.125	1.039 - 1.218

Estimates of model parameters using these values for λ produces results presented in table 3. By comparing these estimates with those presented in table 1, we can see that for all comparisons and in both samples the intercepts decrease and the slopes increase, though these changes are much steeper for the Affymetrix-related comparisons, whereas for Illumina vs Agilent the differences do not appear to be that relevant.

Graphical results are reported in figure 2. Line A498 (upper panels) shows better results with respect to those obtained when $\lambda = 1$, in particular when

Table 3. Estimates of the linear measurement error model, λ estimated

		a_0		b_0	
	Pair	*Estimate*	CI95%	*Estimate*	CI95%
A498	Agilent vs Affymetrix	-7.902	(-7.965, -7.838)	2.048	(2.021, 2.076)
	Illumina vs Affymetrix	-6.497	(-6.556, -6.437)	2.407	(2.380, 2.434)
	Illumina vs Agilent	2.804	(2.545, 3.063)	1.171	(1.104, 1.238)
hREF	Agilent vs Affymetrix	-4.160	(-4.250, -4.071)	1.618	(1.585, 1.651)
	Illumina vs Affymetrix	-0.668	(-0.779, -0.558)	1.720	(1.683, 1.756)
	Illumina vs Agilent	3.757	(3.555, 3.958)	1.062	(1.009, 1.114)

Table 4. Number and proportion of miRNA lying within the agreement intervals, estimated according to the measurement error model parameters estimated by setting $\lambda = 1$ and by estimating it via random effects models. Confidence intervals for the proportions were computed using the Clopper-Pearson exact method [29].

		$\lambda = 1$		λ estimated	
	Comparison	% (*CI*95%)	*n*	% (*CI*95%)	*n*
A498	*Agilent - Affymetrix*	74.17 (71.02 - 77.15)	603	78.84 (75.87 - 81.60)	641
	Illumina - Affymetrix	56.46 (52.97 - 59.90)	459	73.43 (70.25 - 76.44)	597
	Illumina - Agilent	95.45 (93.78 - 96.78)	776	96.31 (94.77 - 97.50)	783
hREF	*Agilent - Affymetrix*	82.53 (79.75 - 85.08)	671	84.26 (81.57 - 86.70)	685
	Illumina - Affymetrix	82.78 (80.01 - 85.31)	673	89.91(87.63 - 91.90)	731
	Illumina - Agilent	97.79 (96.52 - 98.68) †	795	97.54 (96.23 - 98.49) †	793

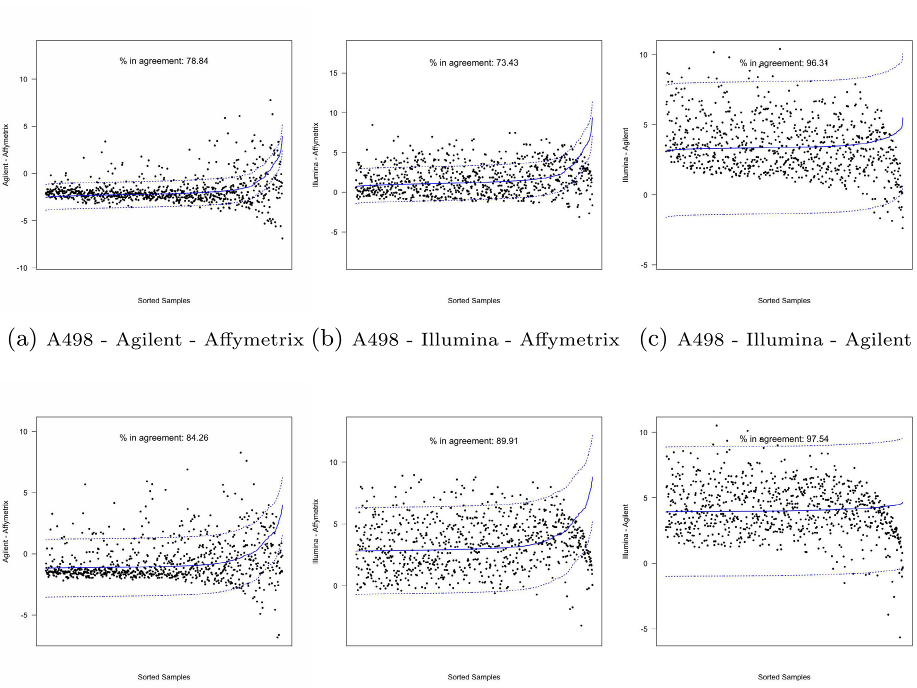
†: the platform pair is in agreement

Affymetrix platform is involved. In fact, an increase of almost 5% is seen when Agilent and Affymetrix are compared: from 74.17% (CI95%: 71.02 - 77.15) with $\lambda = 1$ to 78.84% (75.87 - 81.6) with λ estimated. Additionally, if we consider the Illumina-Affymetrix comparison, there's an increase in the proportion of concordant miRNA from 56.46% (CI95%: 52.97 - 59.90) to 73.43% (CI95%: 70.25 - 76.44). For hREF, Illumina and Agilent show similar patterns of concordance with respect to the value of λ: 97.79% (CI95%: 96.52 - 98.68), i.e. 795 miRNA, for $\lambda = 1$ and 97.54% (CI95%: 96.22 - 98.49), i.e. 793 miRNA, for $\lambda = 1.125$ (estimated value). The comparison Affymetrix - Agilent shows a slight increase in the percentage of concordant miRNA, with only 84.26% (CI95%: 81.57 - 86.70), i.e. 685, miRNA in the agreement interval, whereas Illumina and Affymetrix show the largest increase in proportion of concordant miRNA: 89.91% (CI95%: 87.64 - 91.9).

In table 4 detailed results on percentages of microRNA lying within the agreement intervals visualized in figures 1 and 2 are reported. By comparing results not only according to the parameter λ but also with respect to the different samples, we can note that Affymetrix platform shows moderate-to-low levels of concordance with both Agilent and Illumina irrespectively of the sample considered. In general, the worst comparison appears to be the one between Affymetrix and Illumina, which shows a relevantly lower proportion of concordant miRNAs for all λ and samples.

To have a better view of these results, we have plotted the point and interval estimates separately for each sample in figure 3, but keeping the same range for the Y-axis, to better highlight differences. Let us focus on hREF results, since the technical issues on A498 lines may lead to biased conclusions: Illumina and Agilent show similar patterns of moderate-to-poor agreement with Affymetrix platform when $\lambda = 1$ is considered, but when we estimate it the number of concordant miRNAs seems to increase slightly more for the Illumina-Affymetrix than for the Agilent-Affymetrix comparison. Notably, Illumina and Agilent show very satisfactory levels of agreement in both samples for both choices for the parameter λ.

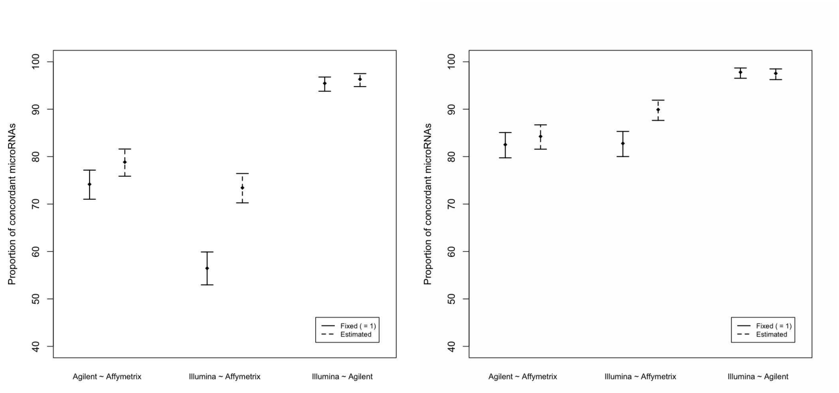
Finally, to assess agreement we need to choose a value for the threshold k, the maximum number of subjects (i.e. microRNA) that we are willing to accept



(a) A498 - Agilent - Affymetrix (b) A498 - Illumina - Affymetrix (c) A498 - Illumina - Agilent

(d) hREF - Agilent - Affymetrix (e) hREF - Illumina - Affymetrix (f) hREF - Illumina - Agilent

Fig. 2. Agreement intervals for the modified Bland-Altman method, λ estimated



(a) A498 (b) hREF

Fig. 3. Confidence intervals for the proportion of concordant miRNAs according to different choices/estimates for the parameter λ

to lie outside the estimated agreement interval. By choosing a value of k =41 ($\simeq$0.05 *813), we can say that only Illumina and Agilent can be considered to be in agreement, whereas Affymetrix shows moderate-to-poor concordance with both Agilent and Illumina.

4 Discussion

Assessing agreement between different methods of measurement is a task which has rarely been addressed in microarray literature, where the focus has always been more on the evaluation of a linear relationship between platforms or with "gold-standard" assays such as qPCR, mainly via computation of correlation coefficients. Such an approach has often been biased by the selection of miRNAs used to do the computations, in that often only miRNAs concordantly detectable between platforms [12, 11] were chosen, possibly leading to an overestimate of the real level of correlation between arrays. To avoid this issue we have considered all common human microRNA that were common (i.e. matched by name) on all the platforms considered for the experiment. Additionally, correlation coefficients are computed assuming that intensity/expression values do not suffer from some measurement error, thus leading to possible under-estimates of the real level of correlation between platforms [15], whereas the proposed method takes into account the presence of an additive measurement error in the measurements from both platforms being compared and incorporates it in the linear model to estimate relationship between arrays.

The method is a modified version of the well-known Bland-Altman plot, a graphical technique commonly used to assess agreement between clinical measurements, and we have applied it in the field of microRNA array platforms by using it in a slightly different way. In a nutshell, commonly there are n subjects on which we measure some biological quantities using k measurement methods ($k \geq 2$) and our goal is to evaluate whether measurements from the k methods agree using information on n samples. Had we followed this procedure, each miRNA should have been evaluated separately (since the microRNA is the biological quantity of interest) and the intensities in the n samples (in our case the two cell lines) would have been compared between the k platforms (in our case 3) jointly for both cell lines. Actually, we have considered the microRNAs to be "subjects" (so that $n = 813$) and the whole profile of intensity on a platform to be the vector of measurements to be compared between different platforms (that is, $k = 3$), separately for each cell line.

By setting a "strong" threshold at 95%, equivalent to 772 microRNAs, we conclude that Agilent and Illumina arrays are concordant according to agreement interval derived from the estimation of the measurement error model, irrespective of the method for choosing the value of λ. The choice of this threshold is subjective, and should depend on the issue at hand; our choice was due to the fact that it is likely that a few miRNAs exist which do not agree because of unwanted technical issues not attributable to the platform itself, but also to the need to reduce the number of false positives (i.e. falsely concordant platforms). This last point is crucial in the field of microarrays, since the comparability of results from different platforms as a tool for validating a lab's own results has gained much relevance in omic research and relies on the assumption that the two platforms are "linked" in that what they say about the profile of intensity/expression for a sample is similar, net of the different measurement and analytic scale of the platform itself.

One of the major issues here was, of course, the choice/estimation of the λ parameter. We have already discussed that setting it equal to 1 is a strong assumption which, though reasonable in many practical situations, is likely to be violated in microarrays. To circumvent this issue, we estimated the error variance by means of replication, fitting a random effects linear model to each platform and sample, but also this strategy presents some drawbacks, since the assumptions underlying the model can be questionable. In the context of random effects models, a Bayesian framework exists that might result in better estimates, yet its performance need to be carefully studied [30].

Notably, the measurement error model considered here is just one of the possible models that could be fitted. Actually, the linearity of the relationship between poorly concordant platform pairs can be questioned, so that the differences we have found could be due to a non-linearity of the relationship or to a lack-of-fit of the regression line, which can be accounted for by considering different functional forms for the x variable, both in a linear and in a non-linear context. In particular, non linear measurement error models [31] and regression splines [32] can be considered as valuable alternatives.

5 Conclusions

In this study we have addressed the issue of between-platform agreement between three different miRNA microarray platforms by making use of a modified version of the Bland-Altman plot which incorporates a measurement error linear regression model to build corrected agreement intervals. Our results show that Agilent and Illumina were the most concordant platform showing good patterns of agreement, whereas Affymetrix-related comparisons showed poor agreement for both lines. Whereas for line A498 this could be explained by technical issues on one replicate, for line hREF this is possibly due to a non-linear relationship between the arrays, thus suggesting to consider different functional forms to achieve a better characterization of the relationship.

The proposed method can thus be used to assess agreement in various contexts and, though supposing a simple linear relation exists between arrays, allows to estimate it in terms of model parameters corrected for the presence of measurement error, an issue which is often neglected in microarray studies.

References

1. Esquela-Kerscher, A., Slack, F.J.: Oncomirs - microRNAs with a role in cancer. Nature Reviews Cancer 6(4), 259–269 (2006)
2. Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R., Golub, T.R.: MicroRNA expression profiles classify human cancers. Nature 435(7043), 834–838 (2005)
3. Blower, P.E., Verducci, J.S., Lin, S., Zhou, J., Chung, J., Dai, Z., Liu, C., Reinhold, W., Lorenzi, P.L., Kaldjian, E.P., Croce, C.M., Weinstein, J.N., Sadee, W.: MicroRNA expression profiles for the NCI-60 cancer cell panel. Molecular Cancer Therapeutics 6(5), 1483–1491 (2007)
4. Søkilde, R., Kaczkowski, B., Podolska, A., Cirera, S., Gorodkin, J., Møller, S., Litman, T.: Global microRNA analysis of the NCI-60 cancer cell panel. Molecular Cancer Therapeutics 10(3), 375–384 (2011)
5. Hébert, S.S., Horré, K., Nicolaï, S., Bergmans, B., Papadopoulou, A.S., Delacourte, A., De Strooper, B.: MicroRNA regulation of Alzheimer's Amyloid precursor protein expression. Neurobiology of Disease 33(3), 422–428 (2009)
6. Dehwah, M.A., Huang, Q.: MicroRNAs and Type 2 Diabetes/Obesity. Journal of Genetics and Genomics 39(1), 11–18 (2012)
7. Paraboschi, E.M., Soldà, G., Gemmati, D., Orioli, E., Zeri, G., Benedetti, M.D., Salviati, A., Barizzone, N., Leone, M., Duga, S., Asselta, R.: Genetic Association and Altered Gene Expression of Mir-155 in Multiple Sclerosis Patients. International Journal of Molecular Sciences 12(12), 8695–8712 (2011)
8. Ach, R.A., Wang, H., Curry, B.: Measuring microRNAs: Comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. BMC Biotechnology 8, 69 (2008)
9. Miska, E.A., Alvarez-Saavedra, E., Townsend, M., Yoshii, A., Sestan, N., Rakic, P., Constantine-Paton, M., Horvitz, H.R.: Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biology 5(9), R68.1–R68.13 (2004)
10. Yauk, C.L., Berndt, M.L.: Review of the Literature Examining the Correlation Among DNA Microarray Technologies. Environmental and Molecular Mutagenesis 48(5), 380–394 (2007)
11. Sato, F., Tsuchiya, S., Terasawa, K., Tsujimoto, G.: Intra-Platform Repeatability and Inter-Platform Comparability of MicroRNA Microarray Technology. PLoS ONE 4(5), e5540 (2009)
12. Yauk, C.L., Rowan-Carroll, A., Stead, J.D.H., Williams, A.: Cross-platform analysis of global microRNA expression technologies. BMC Genomics 11, 330 (2010)
13. Chen, J.J., Hsueh, H.M., Delongchamp, R.R., Lin, C.J., Tsai, C.A.: Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC Bioinformatics 8, 412 (2007)
14. Bland, J.M., Altman, D.G.: Measurement error and correlation coefficients. British Medical Journal 313(7048), 41–42 (1996)
15. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327(8476), 307–310 (1986)
16. Altman, D.G., Bland, J.M.: Measurement in medicine: the analysis of method comparison studies. Statistician 32, 307–317 (1983)
17. http://www.mirbase.org/

18. Liao, J.J.Z., Capen, R.: An Improved Bland-Altman Method for Concordance Assessment. The International Journal of Biostatistics 7(1), 9 (2011)
19. Affymetrix: Affymetrix© miRNA QC Tool guide, Santa Clara, California (2008)
20. Lopez-Romero, P., Gonzales, M.A., Callejas, S., Dopazo, A., Irizarry, R.A.: Processing of Agilent microRNA array data. BMC Research Note 3(18) (2010)
21. R Development Core Team. R: A Language and Environment for Statistical Computing, Vienna, Austria (2011) ISBN 3-900051-07-0, `http://www.R-project.org/`
22. Lopez-Romero, P.: Pre-processing and differential expression analysis of Agilent microRNA arrays using the *AgiMicroRna* Bioconductor library. BMC Genomics 12, 64 (2011)
23. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J.: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80 (2004)
24. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2), 249–264 (2003)
25. Giard, D.J., Aaronson, S.A., Todaro, G.J., Arnstein, P., Kersey, J.H., Dosik, H., Parks, W.P.: In vitro cultivation of human tumors: establishment of cell lines derived from a series of solid tumors. J. Natl. Cancer Inst. 51(5), 1417–1423 (1973)
26. Hua, Y.J., Tu, K., Tang, Z.Y., Li, Y.X., Xiao, H.S.: Comparison of normalization methods with microRNA microarray. Genomics 92, 122–128 (2008)
27. Rao, Y., Lee, Y., Jarjoura, D.: A Comparison of Normalization Techniques for MicroRNA Microarray Data. Statistical Applications in Molecular Genetics Biology 7, 22 (2008)
28. Pradervand, S., Weber, J., Thomas, J., Bueno, M., Wirapati, P.A., Lefort, K., Dotto, G.P., Harshman, K.: Impact of normalization on miRNA microarray expression profiling. RNA 15, 493–501 (2009)
29. Clopper, C., Pearson, E.S.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413 (1934)
30. Chung, Y., Rabe-Hesketh, S., Gelman, A., Liu, J., Dorie, V.: Avoiding Boundary Estimates in Linear Mixed Models Through Weakly Informative Priors. U.C. Berkeley Division of Biostatistics Working Paper Series, paper 284 (2012)
31. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models: A Modern Perspective. Chapman & Hall/CRC, New York (2006)
32. Berry, S.M., Carroll, R.J., Ruppert, D.: Bayesian Smoothing and Regression Splines for Measurement Error Problems. Journal of the American Statistical Association 97(457), 160–168 (2002)

Extracting Key Pathways from Gene Signature and Genetic Aberrations in Subtypes of Cancer

Peikai Chen[1], Yubo Fan[2], Tsz-kwong Man[3], Ching C. Lau[3],
Y.S. Hung[1], and Stephen T.-C. Wong[2]

[1] Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong
{pkchen,yshung}@eee.hku.hk
[2] Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, USA
{yfan,stwong}@tmhs.org
[3] Texas Children's Cancer and Hematology Centers and Dan L. Duncan Cancer Center, and Department of Pediatrics, Baylor College of Medicine, Houston, USA
{ctman,cclau}@txch.org

Abstract. Subtypes of cancer are characterized with subtype-specific aberrations and gene signature. While the gene signature is related to the consequences of the cancerous process, some of the genetic abnormalities such as copy number aberrations (CNAs) can have tumorigenic roles by perturbing various biological pathways. Bridging the gap between the aberrations and signature genes, by extracting networks that reflect the within-subtype variations, may help gain insights on the mechanisms of a cancer and its subtypes.

We report a systemic approach to extract pathways. Using multivariate regression, we model the expression of a signature gene as dependent on the CNA-affected genes. The weighted ℓ_1-norm penalty on the regression produces a sparse matrix, from which a bipartite graph is extracted and subtype specific networks uncovered. For each individual network, we develop an network-growing algorithm by utilizing within-subtype variations, to further identify non-signature targets. To evaluate the clinical relevance of the extracted networks, we derived a goodness-of-fit metric based on Cox proportional hazard rate model and ranked the networks based on this metric. The method was applied to two medulloblastoma datasets and the resulting networks demonstrate both dataset-invariance and biological-interpretability.

Keywords: pathways, copy number aberrations, cancer, subtypes, LASSO, ℓ_1-norm.

1 Introduction

Pathways play important roles in the normal functioning of biological systems, and when dysregulated, the development of cancer as well. Pathways can be dysregulated by a number of factors, including aberrant genetic events such as copy

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 132–146, 2013.

number aberrations (CNAs), causing loss-of-function or gain-of-function consequences that may be tumorigenic. In fact, genetic abnormalities are important features of the cancer genome [1]. However, given the massive numbers of observable and un-observable *errors* in the cancer genome, detecting and establishing the cancer-causing ones is challenging.

On the other hand, a cancer is now believed to consist of several major subtypes, each with distinct behaviors ranging from molecular profile, clinical performance to drug response, etc. These cancer subtypes are now found in breast cancer [2], glioblastoma [3] and medulloblastoma [4], etc. Some theories of cancer subtypes suggest that they may have been caused by the hits (e.g. CNAs) on different pathways, such as Wnt and Shh pathways [5]. Therefore, uncovering the pathways in the subtypes would help understand the underlying mechanisms.

Subtype non-specific pathway analysis is widely studied. For example, Lee *et al.* [6] developed an algorithm that, given a pathway, greedily searches a subset of its member genes that demonstrate significant condition-responsive expressions. A score based on the weighted sum of the t-statistics of individual genes in this subset is assigned to the pathway as an activity indicator. Hundreds of curated pathways are tested before some top ones surface as disease-related candidates. More works in oncogenic pathway identification can be found in [7–10].

By and large, it came to be realized that these disease related expression networks may be the consequences of certain perturbations at the DNA level. Perturbing factors including mutations [11], single-nucleotide polymorphisms (SNPs) [12], CNAs [13], microRNAs [14], etc. were tested on various gene expression datasets to identify the error-induced networks. Most of these approaches make use of pairwise regression between the regulators and the expression networks. But in a real network, a responsive gene could be the superimposed effect of regulations by multiple regulators; conversely, a regulator can regulate multiple targets. As a result, individual association between pairs of genes may overstate the co-regulations. These issues will have to be addressed.

To this end, a systemic approach to extract genetic aberrations perturbed network in cancer subtypes is proposed which shall be elaborated in the sequel of the paper. We applied the approach to the pediatric brain tumor of medulloblastoma, and report results thereof, in later sections.

2 Approach

Suppose we are given a cancer dataset with n samples, the i-th member of which contains a measurement vector $(\mathbf{x}_i, y_i, \mathbf{c}_i)$, where $\mathbf{x}_i \in \mathbb{R}^M$ is the set of expressions for all M genes, $y_i \in \{1, ..., K\}$ is the class label, $\mathbf{c}_i = \{c_{ij} | c_{ij} \in \{0, 1, 2, 3, ...\}, j = 1, ..., M\}$ is the copy number vector. Here, the class labels are obtained by the subtyping method described in Section 5.1. For each subtype k, a set of signature genes $\Omega_k \in \{1, ..., M\}$ are obtained by the method described in Section 5.3; and a set of recurrent CNA-affected genes specific to k, denoted as $\Phi_k \in \{1, ..., M\}$ are detected by the method in Section 5.2. The copy number c_{ij} is discrete and $c_{ij} = 2$ refers to the normal case, otherwise it is either losses

($c_{ij} = 0, 1$) or gains ($c_{ij} > 2$). The following assumes we are building networks for subtype k only, but the process is applicable to all subtypes; and for simplicity, denote Ω_k as Ω, Φ_k as Φ, and $N_\Omega = |\Omega|$ and $N_\Phi = |\Phi|$.

2.1 Weighted-Penalized Regression for Identifying CNA Regulators

For a signature gene $\omega \in \Omega$, whose expression is denoted by a variable X_ω, we want to know the regulators whose expressions X_ω depends on. Further, although we exclude genes with recurrent CNAs to be in the signature (Section 5.2), ω may contain some non-recurrent CNAs among the samples. As a result, X_ω may be affected by its own copy numbers, which we use a variable X_ω to denote. Assuming linear dependency, we have:

$$X_\omega = \mu_\omega + \alpha_\omega C_\omega + \sum_{\varphi \in \Phi} \beta_\varphi X_\varphi + \epsilon_\omega \tag{1}$$

where X_φ is the variable denoting the expression of a candidate regulator $\varphi \in \Phi$ and β_φ the corresponding coefficient, and ϵ_ω the error term assumed to be from identically independently Gaussian distribution $\mathcal{N}(0, \sigma_\epsilon^2)$. C_ω is the copy number profile of gene ω. For simplicity, it is denoted that $\theta = (\mu_\omega, \alpha_\omega, \beta_\varphi)$, and $Z_\omega = (1, C_\omega, X_\varphi)$, so that $X_\omega = \theta^T Z_\omega + \epsilon_\omega$. Unfortunately, the number of candidate regulators N_Φ is often prohibitively huge (in the hundreds, say). Further, only a few top regulators are of interest, whereas regulators with weak inter-dependency with X_ω are assumed to be not functionally related. For this purpose, an ℓ_1-norm penalty is imposed on θ, leading to a log-likelihood function:

$$\mathscr{L}(\theta; \mathbf{x}_\omega, \mathbf{Z}) = -\sum_{i=1}^{n} (x_{i\omega} - \theta^T \mathbf{z}_i)^2 - \lambda ||\theta||_{\ell_1} + \text{Const.} \tag{2}$$

$$= -(\mathbf{x}_\omega - \mathbf{Z}^T \theta)^2 - \lambda ||\theta||_{\ell_1} + \text{Const.} \tag{3}$$

where $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_n]$, $\mathbf{z}_i = [1, c_{i\omega}, \{x_{i\varphi}\}]^T$, *Const.* is a parameter independent constant, and λ is a positive scalar serving as a scale of penalty. Maximization of Eq. 3 is equivalent to LASSO, which constrains that $||\theta||_{\ell_1}$ is smaller than a certain value λ_0 [15]. It can be shown that Eq. 3 is concave and there exists a unique solution $\hat{\theta}$ for a given $\lambda > 0$, even when $N_\Phi > n$. Further, most entries of $\hat{\theta}$ are 0 or close to 0, and the exact number of non-zero elements depends on the value of λ. Specifically, small value of λ leads to fewer 0-valued regression coefficients and large λ has a reverse effect. Depending on λ, the regression can be controlled to be arbitrarily over- or under-fit. To reduced the arbitrariness, λ is chosen such that the goodness-of-fit metric $R_\omega^2 = 1 - \text{RSS}/\sigma_\omega^2$ equals an ω-independent ratio, say, $R_\omega^2 = 0.6$, $\forall \omega$. With this, the penalized regression above can be efficiently trained, resulting in a sparse matrix of coefficients $\hat{\Theta} = [\hat{\theta}_1, ..., \hat{\theta}_{N_\Omega}]$.

Note that in here, the regression is conducted across all samples, regardless of their class labels. To see the reason for doing so, denote the expression of a candidate regulator φ by X_φ and the copy number by C_φ, and we can have a model:

$$X_\varphi = \mu_\varphi + \gamma_\varphi C_\varphi + \pi_\varphi + \epsilon_\varphi \tag{4}$$

where μ_φ is the expected mean, γ_φ is the copy number coefficient, π represents some other regulations on X_φ, and ϵ_φ is the measurement error. Assuming orthogonality, we have:

$$\mathrm{var}(X_\varphi) = \gamma_\varphi^2 \mathrm{var}(C_\varphi) + \mathrm{var}(\pi_\varphi) + \mathrm{var}(\epsilon_\varphi) \tag{5}$$

Note that Φ has already been chosen in a subtype-specific manner, which means C_φ has most of its CNAs confined within k; and C_φ is mostly in aberrant and monotonic (i.e., mostly gains or mostly losses) states in k since it is recurrent. And the dependency of $\mathrm{var}(\pi_\varphi)$ on subtype varies. Consequently, if we want to see the impact of C_φ on X_ω through X_φ, i.e., the variance of X_ω as explained by $\mathrm{var}(C_\varphi)$, the regression needs to include all samples. The regression directly on C_φ is avoided because large numbers of (candidate) regulator genes tend to be physically adjacent and have the same CNA profile, making it non-distinguishable among them. Also, whereas C_φ can be regarded as a step-functioned event over time, i.e., the copy number is static; X_φ may experience some dynamics, from C_φ or from interactions with the target genes, particularly when φ is in some feedback mechanisms. The inter-dependency between X_ω and X_φ captured at the time of measurement may thus to some degree reflects both the static and dynamic regulations of X_ω by C_φ.

In Eq. 3, $\mathbf{Z}\mathbf{Z}^T \triangleq \mathbf{H}$ may be rank-decifient, i.e., $N_\Phi > n$. Then in the original LASSO, when λ_0 is sufficiently large that the norm-ball $||\theta||_{\ell_1} \le \lambda_0$ includes a solution $\overline{\theta}$ to $\mathbf{x}_\omega = \mathbf{Z}^T\theta$; there will be infinite optimal solutions, given by $\overline{\theta} + \mathscr{N}(\mathbf{Z}^T)$, where $\mathscr{N}(\cdot)$ denotes the null space. In the unconstrained LASSO in Eq. 3, if λ_0 is controlled to be small, but there exist a few candidate regulators whose expressions demonstrate strong collinearity, such that the corresponding rows of $\mathbf{H}$ are highly similar, so are the corresponding columns, then the quadratic objective function has very slow gradient along the surface of the constrained norm-ball. As a result, the numerical algorithm tends to randomly select one of them to be non-zero coefficient and set other collinear regulators to be zero (see Fig. 1A).

To handle such cases, we introduce some weights in θ. This tells the algorithm that when *ties* occur, a preference is made to reduce the randomness. To summarize, the problem is re-formulated as:

$$\begin{array}{ll} \text{minimize} & \sum_\xi w_\xi |\theta_\xi| \\ s.t. & 1 - \frac{1}{n}(\mathbf{x}_\omega - \mathbf{Z}^T\theta)^2/\mathrm{var}(\mathbf{x}_\omega) = r \end{array} \tag{6}$$

where $r \in (0,1)$ and $w_\xi > 0$ are pre-specified. This is a quadratically constrained quadratic programming problem and can be uniquely solved to a desirable accuracy. The weights ω_ξ can either be based on prior knowledge, such as bioinformatics data-bases, or based on prediction from normal samples.

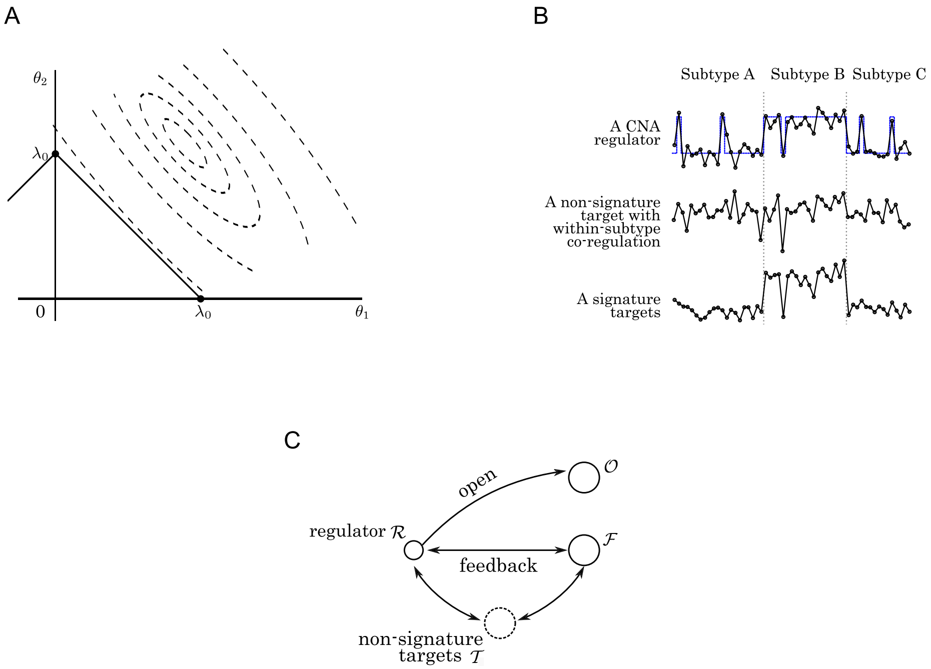


Fig. 1. A, Collinear regulators and the contour of objective function (dashed curves). B, A schematic illustration of within-subtype co-regulations in the non-signature targets. Top-panel: a regulator with copy numbers (blue, predominantly in B) and expressions; middle and bottom: non-signature and signature targets. Horizontal: samples. Note the within-subtype co-expression in Subtype B. C and D, schematic illustration of open-loop and feedback networks.

2.2 Networks Extraction

The sparsity method above results in a bipartite graph from the CNA regulators (i.e., Φ) to the signature genes (i.e., Ω). As a result, multiple signature genes may share a common regulator. If the regulators are functionally responsible for the signature genes, then the co-regulated genes should demonstrate some degree of co-expression.

Note that Ω is also chosen in a subtype-specific manner. Regression in this respect might result in over-optimism as any two signature genes may naturally tend to inter-correlate. To analyze the effect, in a similar fashion as X_φ, we model the variance of X_ω by:

$$\mathrm{var}(X_\omega) = \alpha_\omega^2 \mathrm{var}(C_\omega) + \sum \beta_\varphi^2 \mathrm{var}(X_\varphi) + \sigma_\epsilon^2 \tag{7}$$

$$= \alpha_\omega^2 \mathrm{var}(C_\omega) + \sum \beta_\varphi^2 \gamma_\varphi^2 \mathrm{var}(C_\varphi) + \sum \beta_\varphi^2 \mathrm{var}(\pi_\varphi) + \sum \beta_\varphi^2 \mathrm{var}(\epsilon_\varphi) + \sigma_\epsilon^2 \tag{8}$$

Since within subtype k, $\mathrm{var}(C_\varphi)$ is close to zero, the within-subtype variation of X_ω would largely depend on its own copy number changes and the regulating term $\mathrm{var}(\pi_\varphi)$, plus some error terms. The size of $\mathrm{var}(\pi_\varphi)$ depends on the nature of regulations. If the signature gene is a downstream target, being regulated by an open-loop mechanism, on/off switching effect might dominate, and the within-subtype variations of X_ω might be by pure noise. If instead the signature gene is caught in some feedback processes, the within-subtype variation of X_ω is highly meaningful. Assuming constant noise, the latter case shall have a stronger within-subtype inter-correlation with X_φ. This is illustrated in Fig. 1B.

By the above assumption, the target genes of a regulator can be categorized into two groups: the feed-back ($\mathcal{F}$) and open-loop ($\mathcal{O}$) targets. A framework for network extraction is proposed as below:

1. Perform the penalized regression as in Eq. 6, obtaining a sparse matrix of $\hat{\mathbf{B}} = [\hat{\beta}_1, ..., \hat{\beta}_{N_\Omega}]$. A bipartite graph $G(u, v)$ can be created with $G(u, v) = 1$, if $\hat{\mathbf{B}}_{u,v} > 0$, $\forall$ u and v.
2. Divide the target genes into two groups $\mathcal{F}$ and $\mathcal{O}$. A target gene is assigned to the set $\mathcal{F}$ if its within-subtype correlations with regulator is empirically significant, determined by means of bootstrapping.

After extraction, the network can be summarized by Fig. 1C. It is speculated that some more genes may have been involved in the co-regulation (dashed circles) in-between the regulator set $\mathcal{R}$ and the target sets $\mathcal{F}$ and $\mathcal{O}$. This set of genes is referred to as the non-signature targets $\mathcal{T}$. They may not be significantly differentially expressed as the signature genes as a result of certain feedback mechanisms, but still possess rich clues about the within subtype interactions. Finding this set of non-signature targets may further improve the biological-interpretability of the networks.

Denote the set of non-signature and non-regulator genes by $\Psi = \{\psi | \psi = 1, ..., M, \psi \notin \Omega_k \cup \Phi_k \ , \forall k\}$, and the expression of a gene ψ by a variable X_ψ. Further denote the dependency by D, where $D = 1$ means ψ is dependent and $D = 0$ means not dependent. Then under the two states, the log-likelihood ratio (LLR) is given by:

$$\mathrm{LLR}(\psi) = 2 \log \frac{\sup\{L(D = 1 | X_\psi, X_\imath, X_\jmath)\}}{\sup\{L(D = 0 | X_\psi, X_\imath, X_\jmath)\}} \tag{9}$$

for all $\imath \in \mathcal{F}$ and all $\jmath \in \mathcal{R}$. Here $L(D = 0 | X_\psi, X_\imath, X_\jmath)$ refers to the case where knowing $X_\imath$ and $X_\jmath$ does not increase our knowledge about X_ψ, i.e., $X_\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)$. On the contrary, $L(D = 1 | X_\psi, X_\imath, X_\jmath)$ refers to the case where $X_\psi \sim \mathcal{N}(\mu_\psi + \sum \alpha_\imath X_\imath + \sum \alpha_\jmath X_\jmath, \sigma_\psi^2)$. The statistic $\mathrm{LLR}(\psi)$ can be approximated by a χ_N^2 distribution, where N is the size of the set $\mathcal{F} \cup \mathcal{R}$. To account for the large number of testings, the FWER criterion is imposed.

3 Results and Discussion

We applied the proposed approach to datasets of the pediatric cerebellar tumor of medulloblastoma (MB). Currently, the prognosis for MB patients is poor and

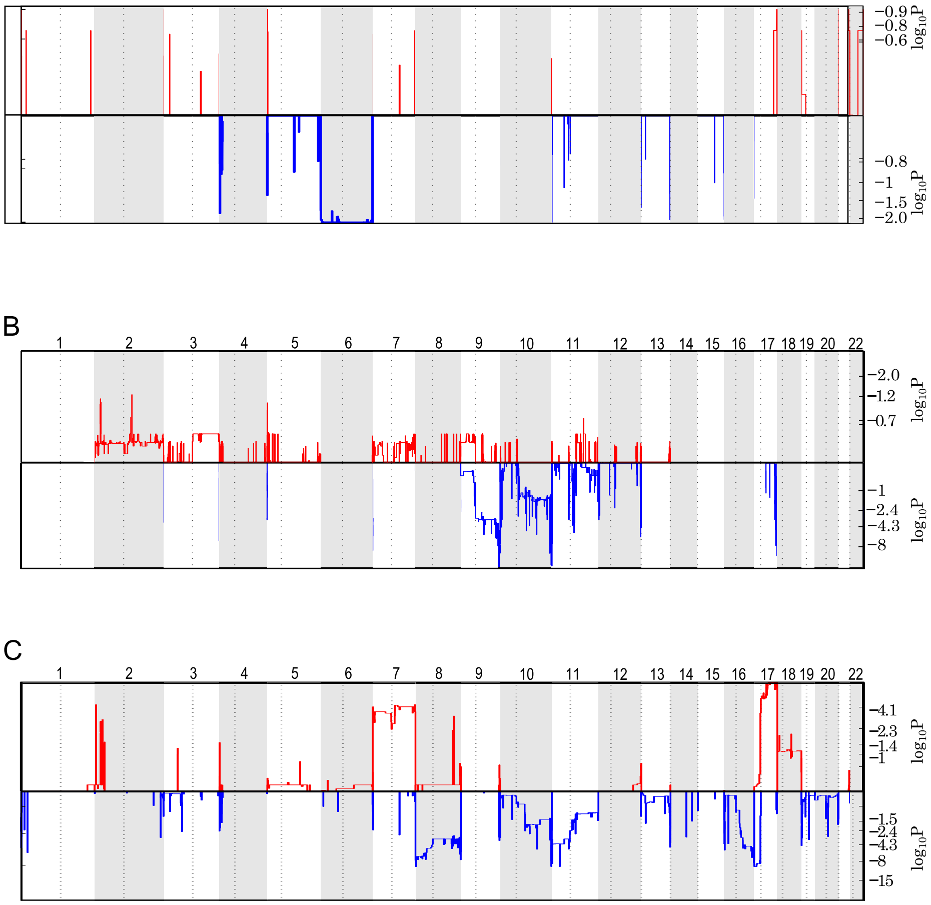


Fig. 2. Subtype-specific copy number aberrations. A, WNT subtype. B, SHH subtype. C: NWS subtype. In each sub-plot, the red curves refer to copy number gains, and blue curves refer to copy number losses.

affected individuals hardly survive to adulthood. Patients of MB demonstrate obvious subgrouping effects in gene expression patterns, copy number profiles, clinical performances, etc. Generally accepted subtypes include the Wnt-pathway and Shh-pathway associated subtypes. Studies have pointed some genetic aberrations, including point mutations [16], loss of heterozygosity [17], copy number aberrations [18], etc., to be susceptible loci. Some of these aberrations occur in signaling pathways critical during brain development. It is of interest to know how aberrations can perturb the normal regulation of these signaling pathways, and indeed, also what pathways are involved other than the well known Wnt and Shh pathways, in the subtypes of MB. The followings present the implementation of the proposed approach and findings thereof.

3.1 Dataset-Invariant and Subtype-Specific Gene Signature and Aberrations

Two publicly available MB datasets, MB1 and MB2 (see 5.1), were used. To facilitate discussion, the three subtypes in MB1 are referred to as WNT, SHH and NWS, respectively; while those in MB2 are referred to as wnt, shh and nws. Gene signature detection by the method 5.3 in were applied to both datasets. Table 1 shows the result. It can be seen that there is strong correspondence between the subtypes in the two datasets. And a substantial percentage of signature genes are dataset-invariant. The overlapping signature genes were used as the set Ω_k for each subtype. This set contains a large number of functionally related genes, that reflect the subtype-specific disease process. For example, in the WNT subtype, some top-ranked genes in Ω_{WNT} include LEF1, WIF1, FZD10, TGFA, WNT11, DKK1, etc., corresponding to the activation of the Wnt-pathway. Likewise, the SHH subtype signature Ω_{SHH} is abundantly enriched with Shh-pathway genes. While the gene signatures in WNT and SHH are quite well-known (and hence the subtype names), a large quantity of novel signature genes were uncovered for the NWS subtype, which is dissected into various numbers of subtypes by recent studies (see 5.1). For example, RREB1, a transcription factor binding to Ras-responsive elements, is uniquely over-expressed in NWS. RREB was shown to be mediating some cancer [19]. Other signature genes include the Wnt-pathway genes NLK, FZD1/7, TCF4, SOX4; Protein Kinase-A pathway genes AKAP6/9, GNG3, H3F3A/B, PLCB4, PP1R10, PRKAR2B, etc. (see Supplementary for a complete list for Ω_{NWS}) To summarize, the signatures of subtypes have strong functional implications about the particular processes underlying the subtypes.

The subtype-specific recurrent CNAs were detected on SNP arrays of MB1 (see 5.2). Fig. 2 shows the subtype-specific copy number landscapes. The WNT subtype is dominant with significant copy number losses on Chr6. The SHH subtype has significantly recurrent losses on Chr9q and Chr10q, and includes genes such as NOTCH1. NOTCH1 is a known suppressor in some cancer [20] and recently shown to interact with Shh-pathway in regulating neocortical progenitors [21]. Copy loss of this gene may reduce its tumor-suppressing capacity. The NWS subtype is characterized with gains on Chr7, Chr17q and losses on Chr8, Chr11 and Chr16. Of note, there is only marginal overlap between the signature genes and CNA-affected genes. For example, $|\Omega_k \cap \Phi_k| = (13, 28, 16)$ for $k =$ (WNT, SHH, NWS), respectively. This perhaps suggests that the signature genes are not mechanic responses induced by their own CNAs, but rather the consequences of some processes, for which a network extraction approach may help bridge the gap.

3.2 Significantly Reproducible Networks Characterize Subtype-Specific Processes

Since our model does not use copy numbers, but instead the expressions of CNA-affected genes. The set of CNA genes can be assumed to be candidate regulators of MB2 as well. In fact, the copy number landscape by arrayCGH in [4] shows

Table 1. Cross-dataset comparison of subtype signatures

	Subtypes of MB2			
Subtypes of MB1	wnt	shh	nws	MB1 signatures
WNT	421	9	25	455 (92.5%)
SHH	13	160	6	323 (49.5%)
NWS	18	0	201	212 (94.8%)
MB2 signatures	568 (74.1%)	194 (82.5%)	310 (64.8%)	

very similar patterns as in MB1. It is found that most of the signature genes have normal copy numbers, the C_ω term in Eq. 1 can be dropped. In this way, both expression datasets can be used to construct the regulatory networks. As a result, the same sets of candidate regulators and signature were used in the two datasets.

We implemented the weighted-penalty ℓ_1-norm regression model to both datasets. The weighting w_φ for a regulator φ was selected based on the within-subtype variation of X_φ. As a result, in case of *ties*, the candidate regulator with higher within-subtype variance (therefore higher $\mathrm{var}(\pi_\varphi)$) is chosen. To estimate the standard errors for the coefficients $\hat{\beta}_\varphi$, a bootstrapping on the samples was performed with 10,000 resamplings. $\hat{\beta}_\varphi$ is deemed significantly non-zero if its 99% empirical confidence interval does not include zero. To determine the non-randomness of the networks, we compute the ratio of dataset-invariant edges, i.e., (u, v) for which $|G(u, v)| > 0$ in both MB1 and MB2, by:

$$\text{ratio} = \frac{\#\text{dataset-invariant edges} * 2}{\#\text{significant edges in MB1} + \#\text{significant edges in MB2}} \tag{10}$$

In all, 11090, 4488 and 6079 significant edges are found in the three subtypes (WNT, SHH and NWS) of MB1, respectively; and another 11289, 4640 and 8034 edges are found in the three subtypes of the MB2 dataset, respectively. Of these, 3829, 1183 and 2099 edges are found in both datasets in the three subtypes, respectively. These correspond to overlapping ratios (Eq. 10) of 34.2%, 25.9% and 29.7%, respectively. These overlapping ratios are all found to be very significant by bootstrapping ($P < 10^{-6}$) and that indicates strong reproducibility of the uncovered networks. These overlapping edges automatically form three networks and are shown in Fig. 3.

3.3 Feedback Genes and Non-signature Targets Capture Within-Subtype Co-regulations

The signature genes are categorized into the feedback genes ($\mathcal{F}$) and open-loop genes ($\mathcal{O}$) by a correlation-based method. A signature gene is said to be in $\mathcal{F}$ if it is highly correlated (among the within-subtype samples) with the regulators, otherwise it is said to be in $\mathcal{O}$. A N_Φ-by-N_Ω matrix Z is formed where each entry $Z_{\varphi,\omega}$ refers to the within-subtype correlation between φ and ω. Suppose Z is row-wise zero-meaned. Take the SVD, $Z = U\Sigma V^T$, and the projection $z_1 = U_{\cdot 1}^T Z$,

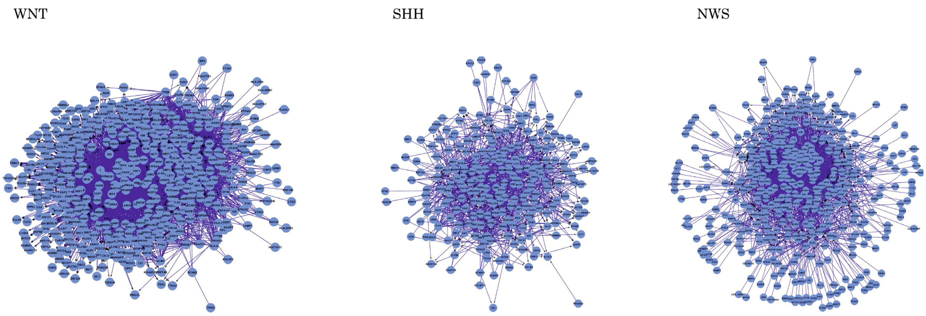


Fig. 3. The reproducible networks pointing from the CNA regulators to the gene signature for each subtype

Table 2. Feed-back genes and non-signature target genes

Subtypes	Feed- back genes ($\mathcal{F}$)	non-signature target genes ($\mathcal{T}$)
WNT	AHI1, CD47, ODZ3, REEP5, PLCB1, IBTK, SMAP1, IFT57, CST6, MAPKBP1, ILVBL, MED23, ACY1, PHIP, KCND3, LOC282997, SFRS18, PELI2, MTO1, TBC1D2B, ZBTB24, FBXO9, MAPKAPK3, AQP3, PRPF4B, COTL1, DSE	KCNE2, AMBP, CXorf57, MED14, GYPA, RHCE, KLHL23, C3orf32, ZNF407, IFNA8, AFF2, HIVEP3, PTN, RPL41, SPINLW1, UPK2, KCNC4, CLTB, PCDHB1, RXRG, DECR2, S100G, SPINT3, TSHB, GNAL
SHH	LPPR4, PDLIM3, GLI1, SLC6A8, GAB1, ABCB4, ATOH1, EPB41L4A, F8A1, KPNA1, EP400, CYLD, LHFPL2, NAP1L1, CBFA2T2, BAHCC1, NAP1L1, ARHGAP19	MUC4, CD8B, KIR3DL1, GPR135, CTDSPL, SSX4, FLJ23519, ACVR1B, CSF2, LOC388907, AQP6, SLC14A2, IGHG1, LOC100131825, OR10H3, ABCB11, IDS, STIM1, GPD2, MFAP3L, TNP2, LDHAL6B, PIP5K1A, PDE4C, KIR2DL1, KCNJ1
NWS	FOXG1, RYBP, H3F3A, TST, CBX1, TES, NRCAM, WIPI1, SOX4, ROR1, KIAA0195, MED13, DYRK2, CDK5R1, SLC6A3, KIDINS220, PGS1, PECR, FZD7	AGTR2, RAB7A, OTC, NEK1, IGJ, OLFM4, IGL@, LTF, HTR3B, MAP2K5, IGH@, MFI2, TFF1, LOC339562, PPYR1, PLD1, GDNF, RNF185, NCKIPSD, SULT1C2, PSG6, NPY2R, SEMG2, PDZRN4, KIR3DX1, CD34, KCNJ15, TNIP3, IL10, SULT2A1, LOC100101117, GLS, IVD, DAZ1, CSF3R, SIX2, HNRNPD, IGHA1, SLC5A1, TACSTD2, IL1R1, THBS1, MFAP5, CPA3, GP5

where $U_{\cdot 1}$ is the first column of U. 1,000 smooth bootstraps are performed to determine the empirical confidence interval (eCI) of $z_1^{(i)}$, for $i = 1, ..., N_\Omega$. If the eCI of $z_1^{(i)}$ does not include zero, it is said to be in $\mathcal{F}$. In total, 28, 19 and 21 genes in the three subtypes are found to be dataset-invariant feedback genes, respectively (see Table 2). Note that some key pathway genes, such as GLI1, ATOH1 that associate with the Shh-pathway appears in the SHH subtype. While few Wnt-pathway genes appear in the WNT subtype. This indicates that the dysregulating mechanism of WNT subtype may be different from that of the SHH subtype, in the sense that the latter contains more within-subtype interaction than the former. This in turn may be due to the multiple feedbacks that fine tune the Shh-pathway [22]. The feedback genes of the NWS are characterized with occasional Wnt-pathway genes such as FZD7 and SOX4, but the roles are not very clear.

The log-likelihood based non-signature target identification procedure (cf. Eq. 9) was applied on each subtype specific network to identify the non-signature target genes ($\mathcal{T}$) and summarized in Table 2. Of note, GSEA analysis reveals that the non-signature target genes of the WNT subtype do not seem to be very relevant with the Wnt pathway. This may confirm the aforementioned theory that the Wnt-pathway may have been permanently turned on in the WNT subtype and as a result feedbacks and within subtype interactions are less obvious. In the SHH subtype, immunity-related pathways are significantly represented, including Antigen processing and presentation (p=0.00453), Natural killer cell mediated cytotoxicity (p=0.0148), etc. In the NWS subtype, there is a significant enrichment of JNK MAPK Pathway (p=0.0238).

4 Conclusion

In this work, we have presented a method, utilizing the orthogonality of variances, to build a penalized model that extract clinically related networks in subtypes of cancer. The results of the method to medulloblastoma demonstrates strong dataset-independence and strong biological interpretability. As an extension, our work indicates that signature genes and aberrations in cancer and its subtypes do not just co-exist by chance, but instead are functionally intercorrelated and experimentally observably. It is thus possible and useful to link them up and uncover the individualized cancer mechanisms. Our successful application to medulloblastoma allows the extension of the approach to other cancers, where subtypes also widely exist.

5 Methods

5.1 Datasets and Subtyping

Two medulloblastoma datasets were used. The first data set containing 74 samples is from Cho *et al.* [23]. The second data set containing 62 samples is from Kool *et al.* [4] (GEO accession number: GSE10327). For convenience, the two data sets are referred to as MB1 and MB2, respectively. Note that MB2 uses a newer gene expression array and contains more probesets than MB1. For simplicity, only probesets common to both MB1 and MB2 are used.

Subtyping results from both the Cho and Kool studies were used as class labels for the corresponding data sets, respectively. Of note, the two studies reported different numbers of subtypes in medulloblastoma. But in both subtyping results, roughly three super-subtypes can be summarized: (1) the Wnt-pathway associated subtype; (2) the Shh-pathway associated subtype, and; (3) the non-Wnt/non-Shh (NWS) patients. While the former two subtypes are now widely accepted, major debates still center around the exact subtyping patterns in the NWS patients. For convenience, we refer to the third category as the NWS subtype. The subtypes and numbers of cases are summarized as below:

current codes	codes in the source	# cases	current codes	codes in the source	# cases
MB1		74	MB2		62
WNT	6	7	wnt	A	9
SHH	3	18	shh	B	15
NWS	1, 2, 4, 5	49	nws	C, D, E	38

5.2 Reccurent CNAs Detection

Only the first dataset, MB1, has public copy number information. The SNP arrays matching with the expressions samples were downloaded from GEO (accession no.: GSE19399). Copy number profiles $\mathbf{c}_i$ for sample i were inferred from the sample-matched SNP arrays, by Genotyping Console (Affymetrix, CA). Regions with significantly recurrent CNAs were detected by GISTIC [24] on GenePattern (genepattern.broadinstitute.org), with subtype-specific data from these SNP arrays. This results in a set of candidate CNA regulators Φ_k for each subtype k. The numbers of candidate regulators in each subtype are summarized as below:

	Φ_{WNT}		Φ_{SHH}		Φ_{NWS}	
	gains	losses	gains	losses	gains	losses
# genes	0	359	0	281	692	561

5.3 Subtype Signature Detection

To identify the subtype signatures, a three step algorithm is developed: (i) detection of differentially expressed genes (DEGs), (ii) detection of subtype-specific DEGs, or subtype signature, and (iii) ranking of genes within a subtype signature.

To detect the DEGs, given expressions $e_j = [x_{1j}, ..., x_{nj}]$ for gene $j \in \{1, ..., M\}$ and the subtype label $\mathbf{y}$, an ANOVA is performed to test: $H_0 : \mu_1^j = .. = \mu_K^j$, where μ_k^j is the mean of the within-subtype mean of gene j for subtype k. To account for the large number of comparisons, we use the family-wise error rate (FWER) as corrected by the Holm-Bonferroni ([25]) method to select the top DEGs.

In a multiclass setting, subtype-specific DEGs can be detected via *post-hoc* analysis. For each of the DEGs detected in the ANOVA, Tukey's ([26]) honest significance test (TukeyHSD) is followed to conduct a pair-wise comparison of its expression in one subtype with that in another. A DEG is said to be specific to a subtype, if the TukeyHSD signs of it in that subtype's comparisons with all other subtypes are identical, i.e., all positive or all negative, and the corresponding adjusted p-values are all significant. It is worth noting that no restrictions are imposed on other adjusted p-values in the TukeyHSD test, which allows for slight inter-subtype variations in the non-specific subtypes.

Finally, a ranking of each subtype's specific genes is needed to provide an order of functional relevance for these genes. To this end, for each subtype, a comparison for each of the detected signature genes in this subtype, against all other samples as a group, is performed. The subtype signature genes are ordered according to their corresponding p-values. LIMMA [27] with BH [28] false discovery rate (FDR) control is applied in this step. At the end of these steps, we obtain a set of ranked signature genes Ω_k for subtype k.

Acknowledgments. This work is partly supported by the HKU UPF and CRCG. CL would like to acknowledge the Simons Foundation. STCW would like to acknowledge NIH grant U54CA149169 and the TT and WF Chao Foundation.

References

1. Stratton, M.R., Campbell, P.J., Futreal, P.A.: The cancer genome. Nature 458(7239), 719–724 (2009)
2. Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O., Botstein, D.: Molecular portraits of human breast tumours. Nature 406(6797), 747–752 (2000)
3. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B.A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H.S., Hodgson, J.G., James, C.D., Sarkaria, J.N., Brennan, C., Kahn, A., Spellman, P.T., Wilson, R.K., Speed, T.P., Gray, J.W., Meyerson, M., Getz, G., Perou, C.M., Hayes, D.N.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17(1), 98–110 (2010)
4. Kool, M., Koster, J., Bunt, J., Hasselt, N.E., Lakeman, A., van Sluis, P., Troost, D., Meeteren, N.S., Caron, H.N., Cloos, J., Mrsic, A., Ylstra, B., Grajkowska, W., Hartmann, W., Pietsch, T., Ellison, D., Clifford, S.C., Versteeg, R.: Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One 3(8), e3088 (2008)
5. Taipale, J., Beachy, P.A.: The Hedgehog and Wnt signalling pathways in cancer. Nature 411(6835), 349–354 (2001)
6. Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T., Lee, D.: Inferring pathway activity toward precise disease classification. PLoS Comput. Biol. 4(11), e1000217 (2008)

7. Ergun, A., Lawrence, C.A., Kohanski, M.A., Brennan, T.A., Collins, J.J.: A network biology approach to prostate cancer. Mol. Syst. Biol. 3, 82 (2007)
8. Chang, J.T., Carvalho, C., Mori, S., Bild, A.H., Gatza, M.L., Wang, Q., Lucas, J.E., Potti, A., Febbo, P.G., West, M., Nevins, J.R.: A genomic strategy to elucidate modules of oncogenic pathway signaling networks. Mol. Cell 34(1), 104–114 (2009)
9. Zhao, J., Gupta, S., Seielstad, M., Liu, J., Thalamuthu, A.: Pathway-based analysis using reduced gene subsets in genome-wide association studies. BMC Bioinformatics 12, 17 (2011)
10. Xu, M., Kao, M.C.J., Nunez-Iglesias, J., Nevins, J.R., West, M., Zhou, X.J.: An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics 9(suppl. 1), S12 (2008)
11. Torkamani, A., Schork, N.J.: Identification of rare cancer driver mutations by network reconstruction. Genome Res. 19(9), 1570–1578 (2009)
12. Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., Leonardson, A., Castellini, L.W., Wang, S., Champy, M.F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T.A., Lusis, A.J., Schadt, E.E.: Variations in DNA elucidate molecular networks that cause disease. Nature 452(7186), 429–435 (2008)
13. Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., Pe'er, D.: An integrated approach to uncover drivers of cancer. Cell 143(6), 1005–1017 (2010)
14. Bonnet, E., Tatari, M., Joshi, A., Michoel, T., Marchal, K., Berx, G., Van de Peer, Y.: Module network inference from a cancer gene expression data set identifies microRNA regulated modules. PLoS One 5(4), e10162 (2010)
15. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
16. Taylor, M.D., Liu, L., Raffel, C., Hui, C.C., Mainprize, T.G., Zhang, X., Agatep, R., Chiappa, S., Gao, L., Lowrance, A., Hao, A., Goldstein, A.M., Stavrou, T., Scherer, S.W., Dura, W.T., Wainwright, B., Squire, J.A., Rutka, J.T., Hogg, D.: Mutations in SUFU predispose to medulloblastoma. Nat. Genet. 31(3), 306–310 (2002)
17. Cogen, P.H., Daneshvar, L., Metzger, A.K., Duyk, G., Edwards, M.S., Sheffield, V.C.: Involvement of multiple chromosome 17p loci in medulloblastoma tumorigenesis. Am. J. Hum. Genet. 50(3), 584–589 (1992)
18. Pfister, S., Remke, M., Benner, A., Mendrzyk, F., Toedt, G., Felsberg, J., Wittmann, A., Devens, F., Gerber, N.U., Joos, S., Kulozik, A., Reifenberger, G., Rutkowski, S., Wiestler, O.D., Radlwimmer, B., Scheurlen, W., Lichter, P., Korshunov, A.: Outcome prediction in pediatric medulloblastoma based on DNA copy-number aberrations of chromosomes 6q and 17q and the MYC and MYCN loci. J. Clin. Oncol. 27(10), 1627–1636 (2009)
19. Thiagalingam, A., De Bustros, A., Borges, M., Jasti, R., Compton, D., Diamond, L., Mabry, M., Ball, D.W., Baylin, S.B., Nelkin, B.D.: RREB-1, a novel zinc finger protein, is involved in the differentiation response to Ras in human medullary thyroid carcinomas. Mol. Cell Biol. 16(10), 5335–5345 (1996)
20. Nicolas, M., Wolfer, A., Raj, K., Kummer, J.A., Mill, P., van Noort, M., Hui, C.C., Clevers, H., Dotto, G.P., Radtke, F.: Notch1 functions as a tumor suppressor in mouse skin. Nat. Genet. 33(3), 416–421 (2003)
21. Dave, R.K., Ellis, T., Toumpas, M.C., Robson, J.P., Julian, E., Adolphe, C., Bartlett, P.F., Cooper, H.M., Reynolds, B.A., Wainwright, B.J.: Sonic hedgehog and notch signaling can cooperate to regulate neurogenic divisions of neocortical progenitors. PLoS One 6(2), e14680 (2011)

22. Varjosalo, M., Taipale, J.: Hedgehog: functions and mechanisms. Genes Dev. 22(18), 2454–2472 (2008)
23. Cho, Y., Tamayo, P., Tsherniak, A., Greulich, H., Lu, J., Kool, M., Zhou, T., Eberhart, C.G., Olson, J.M., Lau, C.C., Meyerson, M., Mesirov, J.P., Pomeroy, S.L.: Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J. Clin. Oncol. 12(6), 1424–1430 (2010)
24. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., Du, J., Kau, T., Thomas, R.K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R.M., Demichelis, F., Hatton, C., Rubin, M.A., Garraway, L.A., Nelson, S.F., Liau, L., Mischel, P.S., Cloughesy, T.F., Meyerson, M., Golub, T.A., Lander, E.S., Mellinghoff, I.K., Sellers, W.R.: Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc. Nat. Acad. of Sci. 104(50), 20007–20012 (2007)
25. Hommel, G.: A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika 75(2), 383–386 (1988)
26. Yandell, B.S.: Practical data analysis for designed experiments. Chapman & Hall texts in statistical science series. Chapman & Hall, London (1997)
27. Smyth, G.K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article 3 (2004)
28. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B-Methodological 57(1), 289–300 (1995)

Investigating Distance Metrics in Semi-supervised Fuzzy c-Means for Breast Cancer Classification

Daphne Teck Ching Lai and Jonathan M. Garibaldi

School of Computer Science, University of Nottingham, UK
{psxdtl,jon.garibaldi}@nottingham.ac.uk

Abstract. In previous work, semi-supervised Fuzzy c-means (ssFCM) was used as an automatic classification technique to classify the Nottingham Tenovus Breast Cancer (NTBC) dataset as no method to do this currently exists. However, the results were poor when compared with semi-manual classification. It is known that the NTBC data is highly non-normal and it was suspected that this affected the poor results. This motivated a further investigation into alternative distance metrics to explore their effect on classification results. Mahalanobis, Euclidean and kernel-based distance metrics were used on 100 sets of randomly-selected labelled data. It was found that ssFCM with Euclidean distance successfully and automatically identified the six classes in close agreement with those of Soria et al. We showed that there is also high agreement in the key features that define the breast cancer classes with those of Soria et al. The superiority of Euclidean distance for classifying this dataset, as compared to Mahalanobis distance is unexpected as it can only generate spherical clusters while Mahalanobis distance can generate hyperellipsoidal ones including spherical ones. We expected Mahalanobis distance to generate the hyperellipsoidal clusters that would best fit NTBC data.

Keywords: semi-supervised, fuzzy c-means, breast cancer classification, distance metrics.

1 Introduction

The Nottingham Tenovus Breast Cancer (NTBC) dataset has been used in studies to understand the mechanism of breast cancer and characteristics of its subgroups [1,2]. The dataset contains 25 immunohistochemical features for 1076 patients. Soria and colleagues [1] successfully identified six clinically novel and useful subgroups while maintaining the three main clinical groups, Luminal, Basal and HER2. However, the methodology used was semi-manual, involving visual inspection and the use of heuristics and other techniques to aggregate results from different unsupervised clustering techniques. A fully automated method (post initialisation) for identifying these subgroups is needed. No unsupervised clustering technique has been found to do this so far.

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 147–157, 2013.

Semi-supervised Fuzzy c-means (ssFCM) is a successful pattern recognition technique applied in many types of biological and medical data. To name a few achievements, Bensaid et al. [3] applied it in MRI image segmentation and Tari et al. [4] grouped functionally related genes with prior knowledge from Gene Ontology annotations. ssFCM have also been shown to produce good classification results in popular benchmark datasets from the UCI repository such as Iris, Wine and Wisconsin Original and Diagnostic Breast Cancer [5,6].

Previously, we applied ssFCM [5] (Pedrycz97) as an automatic method for classifying the dataset with an overarching aim to preserve the main clinical groups and ideally identify the same subgroups as Soria et al. In doing so, we hope to make the prediction of breast cancer types for future patients more efficient. We used ssFCM to perform classification using class labels from [1]. However, a low degree of agreement (Kappa Index) was found.

Distance metrics are an important part of Fuzzy c-means as they are used to measure similarity, which provide additional structural information in terms of the characteristics of data patterns relative to the cluster. The degree of similarity enables us to determine how strongly a data pattern belong to a certain group. Mahalanobis, Euclidean and kernel-based distance metrics use different approach to represent structural information of the dataset, which measure similarity [7]. They are chosen for investigation as they popular distance metrics in ssFCM [3,4,5]. Hidden structural information can be uncovered using suitable distance metrics that can improve classification results.

In the original Pedrycz97 algorithm, Mahalanobis distance metric was used. In this work, we investigate Euclidean and kernel-based distance metrics in Pedrycz97 on varying amounts of labelled data across 100 sets of randomly selected labelled data. Three evaluation techniques, classification rates, Kappa Index and Normalised Mutual Information, are used to measure agreement with classification by Soria et al. because they provide varied treatments to agreeing and disagreeing classification. Clinical insights of NTBC can be found in [1] and are beyond the scope of this paper.

The rest of the paper is organised as follows: The semi-supervised Fuzzy c-means algorithm is explained in Sect. 2, followed by a brief description of the distance metrics used in our study in Sect. 3. In Sect. 4, we describe the dataset, experimental design and evaluation techniques used. Experimental results are reported in Sect. 5 and are discussed in Sect. 6. Finally, the conclusion is found in Sect. 7.

2 Semi-supervised Fuzzy c-Means

Fuzzy c-means is a clustering method which allow a data pattern to belong to more than one cluster, which gives a more realistic representation of data than a binary approach. Membership values are used to indicate the degree of belongingness a data pattern has to clusters and thus determine which cluster a data pattern is assigned to. For a data pattern, membership values to each cluster can range from zero to one and the sum of membership values for all clusters must equal to one. A high membership value to a cluster means a high possibility of belonging to this cluster.

Semi-supervised Fuzzy c-Means (ssFCM) use some labelled data patterns in the dataset to guide the identification of similar data patterns. This can be very valuable when some cases can be labelled. But, labelled data patterns are often sparse and they are time-consuming and labour-intensive to collect.

In [5], the FCM objective function is extended to include supervised learning component represented as the second term as follows:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{p} d_{ik}^{2} + \alpha \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik} - f_{ik} b_k)^p d_{ik}^2, \tag{1}$$

where u_{ik} is the membership value of data pattern k in cluster i, d_{ik} is the distance between data pattern k and cluster centre v_i, f_{ik} is the membership value of labelled data pattern k in cluster i, $\mathbf{b}$ is a boolean vector indicating if a pattern is labelled, c is the number of clusters, N the number of data patterns in the dataset and p is the fuzzifier parameter (which is commonly 2) and α is a parameter for maintaining balance between the supervised and unsupervised learning components.

The algorithm involves iteratively calculating the cluster centres and the membership matrix U containing u_{ik} to minimise the objective function until a termination criterion is satisfied. In this work, we use ssFCM by Pedrycz and Waletzky [5] because it has been shown to produce good classification results. The algorithm is summarised as follows:

1. Initialise labelled data membership matrix $\mathbf{F}$ and initial membership matrix $\mathbf{U^0}$
2. Calculate cluster centres $\mathbf{V} = [\mathbf{v_i}]$ with $\mathbf{U}$ using equation:

$$\mathbf{v_i} = \frac{\sum_{k=1}^{N} u_{ik}^2 \mathbf{x_k}}{\sum_{k=1}^{N} u_{ik}^2} \tag{2}$$

3. Update partition matrix, $\mathbf{U}$ using equation :

$$u_{ij} = \frac{1}{1+\alpha} \left\{ \frac{1 + \alpha(1 - b_j \sum_{l=1}^{c} f_{lj})}{\sum_{l=1}^{c} (\frac{d_{ij}}{d_{lj}})^2} + \alpha f_{ij} b_j \right\} \tag{3}$$

4. If $||\mathbf{U}' - \mathbf{U}|| < \epsilon$, stop. Else, go to step 2 with $\mathbf{U} = \mathbf{U}'$

Note that all data patterns, labelled and unlabelled data patterns undergo unsupervised learning. This means that that memberships of labelled data patterns are updated at each iteration.

3 Distance Metrics

In this section, we briefly describe the distance metrics used. These distance metrics, their differences and behaviours in ssFCM are discussed in detail in [7].

3.1 Mahalanobis

The Mahalanobis distance is formally defined [8] as:

$$d_M(x) = \sqrt{(x = \mu)^T S^{-1}(x - \mu)} \tag{4}$$

It is the distance between a vector $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_n})^\mathbf{T}$ which belong to a group of vectors with a mean $\mu = (\mu_\mathbf{1}, \mu_\mathbf{2}, ...\mu_\mathbf{n})^\mathbf{T}$ and $\mathbf{S}$ is the covariance matrix of the group.

The Mahalanobis distance metric in ssFCM [5] takes into account the membership as well as the similarity between the data pattern and the cluster center. The inverse covariance matrix, $\mathbf{M_i}$ normalises dimensions of different scales, preventing dominance from dimensions with greater scales. Thus, it is scale-invariant. It forms hyperellipsoidal clusters. The Mahalanobis distance is computed as follows:

$$d_M^2(i,k) = (\mathbf{x_k} - \mathbf{v_i})^T \mathbf{M_i}(\mathbf{x_k} - \mathbf{v_i}) \tag{5}$$

where M_i is a positive definite matrix, its inverse defined as:

$$\mathbf{M_i}^{-1} = \left[\frac{1}{\rho_i det(\mathbf{P_i})}\right]^{\frac{1}{n}} \mathbf{P_i} \tag{6}$$

and $\mathbf{P_i}$ is the fuzzy covariance matrices defined as:

$$\mathbf{P_i} = \frac{\sum_{k=1}^{N} u_{ik}^2 (\mathbf{x_k} - \mathbf{v_i})(\mathbf{x_k} - \mathbf{v_i})^T}{\sum_{k=1}^{N} u_{ik}^2} \tag{7}$$

3.2 Euclidean

The Euclidean distance metric forms spherical clusters and does not reflect scale differences among dimensions in high-dimensional datasets. It is computed as follows:

$$d_E^2(i,k) = ||\mathbf{x_k} - \mathbf{v_i}||^2$$

3.3 Kernel-Based

The kernel methods solve non-linear problems by mapping the input space into higher dimensional space (the 'kernel trick' [9]), which is applied to distances metrics in [10]. The idea here is to transform $\mathbf{x_k}$, a data point from a D-dimensional input space to a higher F-dimensional space resulting in $\Phi(\mathbf{x_k})$. The kernel-based distance is defined as:

$$\begin{aligned} d_K^2(i,k) &= ||\Phi(\mathbf{x_k}) - \Phi(\mathbf{v_i})||^2 \\ &= K(\mathbf{x_i}, \mathbf{x_i}) - 2K(\mathbf{x_k}, \mathbf{v_i}) + K(\mathbf{v_i}, \mathbf{v_i}) \end{aligned}$$

We use Gaussian radial basis function as the kernel function in the form:

$$K(a,b) = e^{\frac{-||a-b)||^2}{\sigma^2}}$$

This yield a distance of the form

$$d_K^2(i,k) = 2(1 - K(\mathbf{x_k}, \mathbf{v_i}))$$

Table 1. Number of data patterns in each class and the number of not classified and classified data patterns according to classification by Soria et al

class 1	class 2	class 3	class 4	class 5	class 6	not classified	classified
202	153	80	82	69	77	413	663

4 Experiment

In this section, we describe the dataset and the steps taken to carry out our investigation.

4.1 Dataset

The Nottingham Tenovus Breast Cancer dataset contains 25 immunohistochemical features for 1076 patients. There are three main clinical groups, Luminal, Basal and HER2 and six subgroups where class 1, 2 and 3 belongs to the Luminal group, class 4 and 5 to the Basal group and class 6 to HER2. Each class is described by key features [1] as follows:

- class 1: ER+, PgR+, CK7/8+, CK18+, CK19+, HER3+, HER4+
- class 2: ER+, PgR+, CK7/8+, CK18+, CK19+, HER3-, HER4-
- class 3: ER+, PgR-, CK7/8+, CK18+, CK19+, HER3+, HER4+
- class 4: ER-, p53+, CK5/6+, CK14+
- class 5: ER-, p53-, CK5/6+, CK14+
- class 6: ER-, HER2+

ER and PgR are hormone receptors. CK7/8, CK18 and CK19 are luminal cytokeratins. CK5/6 and CK14 are basal cytokeratins. HER2, HER3 and HER4 are EGFR family members. p53 is a tumour suppressor gene. The + or - at the end of each feature indicates high or low levels respectively. The class distribution and the number of classified and unclassified data patterns are found in Table 1.

4.2 Experimental Design and Set-Up

Figure 1 displays how the experiment is conducted to classify the Nottingham Tenovus Breast Cancer (NTBC) dataset. Labels from classification by Soria et al. are used to generate membership values which are then used to initialise the supervision matrix $\mathbf{F}$ which contains membership values for labelled data. Instead of using random initialisation of membership values, we use the supervision matrix $\mathbf{F}$ to initialise the membership matrix $\mathbf{U^0}$. In doing so, a better starting point is given to the algorithm instead of a random starting point. We use only data patterns which are classified by Soria et al. for investigation and the 413 data patterns which are not classified are disregarded.

We experimented with varying amounts of labelled data; 10%, 20%, 30%, 40%, 50% and 60% of the 663 classified data patterns. To select data patterns

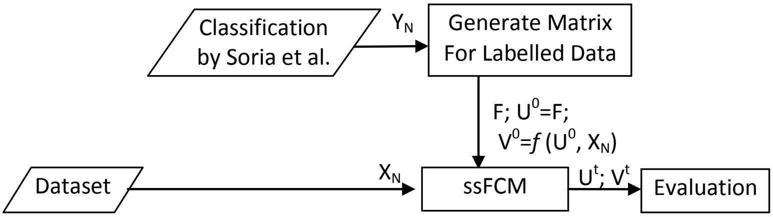

Classification by Soria et al.
Y_N
Generate Matrix For Labelled Data
F; U^0=F; V^0=f (U^0, X_N)
Dataset
X_N
ssFCM
U^t; V^t
Evaluation

Fig. 1. Flowchart of the experiment to classify the Nottingham Tenovus Breast Cancer dataset

to be labelled, an equal number of data patterns is randomly selected from each class. This is to prevent data patterns from a particular class to be selected more frequently than others. We experiment with each varying amount across 100 different sets of labelled data.

From the selected labelled data, we initialise the **F** matrix. To initialise membership values in **F**, the selected labelled data patterns belonging to their respective classes will be given a membership of 0.9 and (1-0.9)/(6-1)=0.02 for classes they do not belong to. The high 0.9 membership value is arbitrarily chosen to indicate a data pattern's high possibility of belonging to the class while a 0.02 value indicates otherwise. Unlabelled data patterns have a membership value of $1/c \approx 0.1667$ to indicate equal possibility of belonging to the classes. In the original Pedrycz97, all data patterns are assigned memberships based on their given labels and stored in **F**. They are then selected to be labelled and unlabelled for the algorithm using the boolean vector in Equation (1), **b**. In our case, we have selected the labelled data for the algorithm and generated their memberships prior to running the algorithm. We set our $\mathbf{F} = \mathbf{U^0}$, where they contain memberships of both labelled and unlabelled data and b_k is 1 for all k (in Equation (3)). The α value is set to be N/M where M is the number of labelled data.

To determine the class of a data pattern $\mathbf{x_k}$, we choose the class with the highest membership value. The classes assigned by ssFCM to the 663 data patterns are then compared with classification by Soria et al. using various evaluation techniques, which will be explained next.

4.3 Evaluation

Three different measures are used to evaluate the performance of the ssFCM algorithm. They are classification rate, Normalised Mutual Information and Cohen's Kappa Index. They are briefly explained as follows:

The Classification Rate (CR) simply calculates the number of matching classification over the total number of data patterns.

Normalised Mutual Information (NMI) [11] calculates the comparison of clusterings in terms of label matching and distribution and normalises this calculation. The NMI equation is as follows:

$$\mathrm{NMI}(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \tag{8}$$

$I(X;Y)$ denotes Mutual Information between variables X and Y and $H(X)$ and $H(Y)$ denote the entropy of variables X and Y respectively. $I(X;Y)$ is computed as follows:

$$I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X) \tag{9}$$

where $H(X|Y)$ and $H(Y|X)$ are conditional entropies and H(X,Y) are joint entropy. NMI values close to zero denotes poor classification while a near 1 value indicates otherwise.

The Cohen's Kappa (κ) Index [12] is given by:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the ratio of agreements between the two sources and p_e is the ratio of chances of agreement.

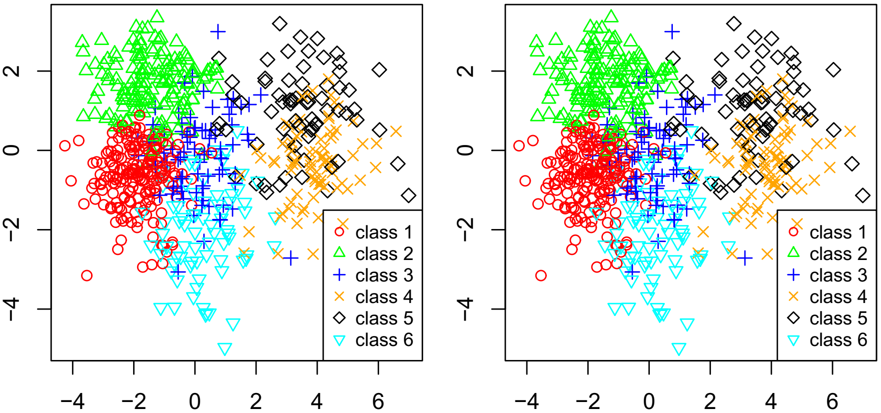
Classification rate tends to give a more optimistic view than Kappa Index and NMI because it only takes into account of agreements and completely disregards disagreements. Both Kappa Index and NMI take into account of agreements and disagreements where there is some sort of penalty for disagreements. In NMI, $H(X|Y)$ and $H(Y|X)$ represent the disagreements. In Kappa, the disagreements are taken into account in the form of a probability of random agreement, p_e.

5 Results

Table 2 shows the classification rates, Kappa and NMI values respectively for classifying NTBC data using the three distance metrics, Mahalanobis (M), Euclidean (E) and kernel-based (K). The results are displayed in the form:

Table 2. Results from evaluation techniques (ET), classification rates (CR), Kappa (κ) and Normalised Mutual Index (NMI) values for NTBC data using distance metrics (DM), Mahalanobis (M), Euclidean (E) and Kernel-based (K) distances. The results from the best performing distance metric is italicised.

DM	ET	10%	20%	30%	40%	50%	60%
M	CR	0.500±0.060	0.587±0.054	0.664±0.044	0.713±0.046	0.757±0.051	0.792±0.055
	κ	0.370±0.063	0.482±0.061	0.575±0.053	0.639±0.057	0.693±0.065	0.737±0.069
	NMI	0.363±0.032	0.436±0.030	0.499±0.030	0.553±0.033	0.604±0.041	0.649±0.049
E	CR	*0.819±0.028*	*0.896±0.025*	*0.941±0.015*	*0.970±0.007*	*0.983±0.005*	*0.991±0.003*
	κ	*0.775±0.035*	*0.871±0.031*	*0.927±0.019*	*0.963±0.008*	*0.979±0.006*	*0.989±0.004*
	NMI	*0.767±0.022*	*0.827±0.019*	*0.879±0.018*	*0.926±0.013*	*0.952±0.012*	*0.972±0.010*
K	CR	0.250±0.088	0.305±0.062	0.357±0.040	0.437±0.033	0.520±0.022	0.613±0.016
	κ	0.081±0.097	0.154±0.068	0.227±0.047	0.326±0.038	0.422±0.027	0.529±0.020
	NMI	0.402±0.029	0.373±0.028	0.353±0.030	0.365±0.028	0.400±0.025	0.450±0.030



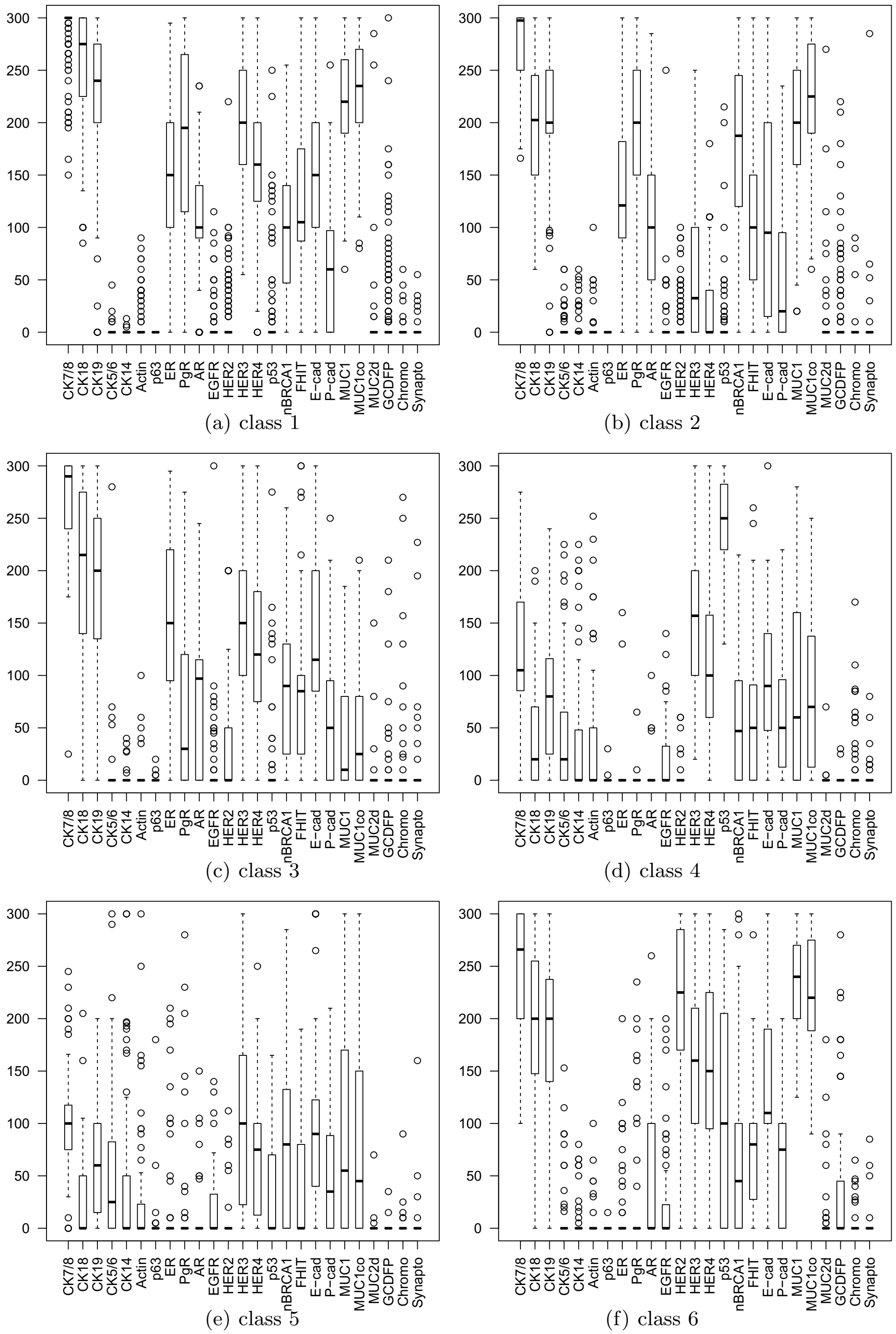
(a) (b)

Fig. 2. Classes of data patterns plotted on a graph using the first and second principal components: (a) Classification from rules by Soria et al. and (b) Classification using Euclidean distance in Pedrycz97 with 60% labelled data

mean±standard deviation. The mean and standard deviation take into account the results obtained from classification with 100 different sets of labelled data. The results show using Euclidean distance produced the best agreements. At only 10% labelled data, a mean Kappa value of 0.775 can be achieved using Euclidean distance, which is very a favourable result. At about 30% labelled data, above 0.9 values can be achieved according to classification rate and Kappa Index. At 60%, it could achieve values near one for all three evaluation measures. The kernel-based distance gave the worst results because even with 60% labelled data, the evaluation records lower values than Mahalanobis and Euclidean distance.

Figure 2 shows a graphical representation of classification by Soria et al. and the best classification using Euclidean distance in Pedrycz97 with 60% labelled data. The best classification in this case is the solution with the highest classification rate, Kappa and NMI values. Principal Component Analysis is used on the dataset for the sole purpose of visualisation and no feature selection has been carried out. The first two components are used to plot the location of the data patterns. It can be observed that the location of the classes are similar in the two classifications, showing that Pedrycz97 with Euclidean distance is able to identify the six subgroups, as those from Soria et al.

Figure 3 show the statistical summary of each features for each of the six classes in the NTBC dataset. Visual comparison of key features which describe classes (found in Fig. 5 in [1]) between this figure and Fig. 4 in [1] reveals high agreement between the two. For instance in class 1, there is agreement in the interquartile range and averages for key features ER, PgR, CK7/8, CK18, CK19, HER3 and HER4 between Fig. 3 and Fig. 4 in [1].



(a) class 1

(b) class 2

(c) class 3

(d) class 4

(e) class 5

(f) class 6

Fig. 3. Boxplot showing statistical summaries of all the features for the six classes obtained from Euclidean distance Pedrycz97 with 60% labelled data

6 Discussion

Euclidean distance in Pedrycz97 successfully classify the NTBC dataset with Kappa values of 0.775±0.035 using 10% and almost complete agreement with a near one Kappa value at 60% of labelled data. The three evaluations give an average of 0.787 at 10% and of 0.984 at 60%, which are very favourable results. Visual inspection of the graphs in Fig. 2 and boxplots comparisons showed that the six classes identified using Pedrycz97 have high agreement with those by Soria et al. Using some supervision information from classification by Soria et al., Pedrycz97 with Euclidean distance is able to automatically produce classification almost identical to that by Soria et al. This also further verify the breast cancer classification by Soria et al.

It is unexpected that the best classification results are produced using Euclidean distance when Mahalanobis have been shown to produce better classifications [6,7]. In [7], Bouchachia and Pedrycz showed that Mahalanobis distance produced the best classification in comparison with Euclidean on UCI datasets Wine, Diabetes and Breast Cancer. Euclidean distance can only generate spherical clusters while Mahalanobis distance can generate hyperellipsoidal clusters, including elliptical and spherical clusters. We are unsure as to why Euclidean distance has produced the most favourable result. Perhaps, the spherical clusters generated by Euclidean distance ideally capture the structural characteristics of the dataset and its subgroups because the features in NTBC are in scale with one another. Perhaps, the non normally-distributed features in NTBC may have been responsible for the unexpected experimental results. It could also perhaps be due to Mahalanobis distance used in ssFCM where the covariance is weighted by fuzzy memberships. The weighted covariance in the distance metric may not necessarily produce a meaningful model. Further investigation is required to ascertain whether the data distribution, the scales of the features or the weighted version of Mahalanobis distance is the reason for the results obtained.

7 Conclusion

Mahalanobis, Euclidean and kernel-based distance metrics are experimented in semi-supervised Fuzzy c-Means algorithm, Pedrycz97 on Nottingham Tenovus Breast Cancer dataset across different sets of labelled data. Using Euclidean distance, Pedrycz97 successfully and automatically classify the NTBC dataset into the six subgroups by Soria et al., further affirming the findings by Soria et al. It was able to achieve an almost complete agreement with classification by Soria et al. This is unexpected as Mahalanobis distance metric has been established to perform better classifications. There are several speculations which require further investigation. Firstly, an investigation in other distance metrics within the Pedrycz97 algorithm to achieve a better classification of the NTBC dataset is to be carried out. Secondly, we wish to look into establishing a relationship between the distance metrics in ssFCM and the nature of the dataset in terms of distribution and scales. Also, further studies is required to find out if a fuzzy weighted version of Mahalanobis distance metric is causing less meaningful model to be produced for this dataset.

References

1. Soria, D., Garibaldi, J.M., Ambrogi, F., Green, A.R., Powe, D., Rakha, E., Macmillan, R.D., Blamey, R.W., Ball, G., Lisboa, P.J., Etchells, T.A., Boracchi, P., Biganzoli, E., Ellis, I.O.: A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. Computers in Biology and Medicine 40(3), 318–330 (2010)
2. Biganzoli, E., Coradini, D., Ambrogi, F., Garibaldi, J., Lisboa, P., Soria, D., Green, A., Pedriali, M., Piantelli, M., Querzoli, P., Demicheli, R., Boracchi, P., Nenci, I., Ellis, I., Alberti, S.: p53 status identifies two subgroups of triple-negative breast cancers with distinct biological features. Japanese Journal of Clinical Oncology 41(2), 172–179 (2011)
3. Bensaid, A.M., Hall, L.O., Bezdek, J.C., Clarke, L.P.: Partially supervised clustering for image segmentation. Pattern Recognition 29(5), 859–871 (1996)
4. Tari, L., Baral, C., Kim, S.: Fuzzy c-means clustering with prior biological knowledge. Journal of Biomedical Informatics 42(1), 74–81 (2009)
5. Pedrycz, W., Waletzky, J.: Fuzzy clustering with partial supervision. IEEE Transactions on Systems, Man and Cybernetics 27(5), 787–795 (1997)
6. Lai, D.T.C., Garibaldi, J.M.: A comparison of distance-based semi-supervised fuzzy c-means clustering algorithms. In: 2011 IEEE International Conference on Fuzzy Systems, pp. 1580–1586 (June 2011)
7. Bouchachia, A., Pedrycz, W.: Enhancement of fuzzy clustering by mechanisms of partial supervision. Fuzzy Sets and Systems 157(13), 1733–1759 (2006)
8. Maesschalck, R.D., Jouan-Rimbaud, D., Massart, D.: The mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 50(1), 1–18 (2000)
9. Aizerman, A., Braverman, E.M., Rozoner, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837 (1964)
10. Schölkopf, B.: The kernel trick for distances. In: Advances in Neural Information Processing Systems, pp. 301–307 (2001)
11. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617 (2003)
12. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)

Two Covariances Harnessing Fuzzy Clustering Based PCA for Discrimination of Microarray Data

Mika Sato-Ilic

Faculty of Engineering, Information and Systems, University of Tsukuba,
Tsukuba, Japan
mika@risk.tsukuba.ac.jp

Abstract. In this paper, we present two kinds of covariance to apply principal component analysis (PCA) to high dimension low sample-size (HDLSS) data. A typical example of this type of data is microarray data. High dimension low sample-size data means data in which the number of dimensions (variables) is much larger than the number of objects (sample size). PCA is a well-known method to reduce the number of variables and to obtain the latent structure of data as the similarity of objects in lower dimensional space spanned by the obtained principal components. It is well known that we cannot obtain correct solutions as the eigen-values of the covariance matrix of variables when dealing with HDLSS data and that the result of ordinary PCA is based on eigen-values of the covariance matrix of variables, therefore if we apply ordinary PCA for HDLSS data, we cannot obtain the correct result. In order to solve this problem, we propose two new types of covariance that exploit the results of fuzzy clustering. First, covariance defines a degree of contribution of objects to clustering as the weights of the covariance by the use of the result of fuzzy clustering. Second we use the correlation between classification structures of two objects which are obtained as the fuzzy clustering results.

Keywords: clustering, fuzzy logic, symbolic data, interval-valued data.

1 Introduction

High dimension low sample-size data is one major concern of today's data analysis. The main reason for this problem is that this type of data has irrelevant and redundant variables related with the curse of dimensionality, so clustering techniques tend to obtain poor clustering results and PCA has a mathematical problem. It is known that the covariance matrix of objects approximately becomes a scaled identity matrix as the number of variables increases with a fixed number of objects, so, all the eigen-values of the covariance matrix of variables are approximately the same in this case. In addition, it is known that the largest eigen-value of sample covariance matrix of variables does not converge to the population counterpart. [1], [2], [6], [8] This means that mathematically we cannot obtain correct solutions as eigen-values of the covariance matrix of variables

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 158–172, 2013.

for HDLSS data. Therefore, if we apply ordinary PCA to HDLSS data in which the number of variables is much larger than the number of objects, then we cannot obtain any correct results. In order to overcome these problems, several clustering methods which use the idea of subsets of variables and the weights of variables for clusters have been proposed. [5], [7] However, in these methods, statistical assumption is needed to define the weights and the calculation to obtain the estimate of the results is complicated. In this paper, we propose simple techniques to obtain the results exploiting the merit of fuzzy clustering in which we can obtain a robust result as continuous values. In addition, we use interval-valued representation which is a type of data representation in symbolic data analysis. [3], [4] Numerical examples show better performance for our proposed methods. In this paper, we present two kinds of principal component analyses for HDLSS data based on two kinds of covariance of variables.

The rest of this paper is structured as follows. Section 2 describes a new covariance of variables including correlation of classification structures. We show that this covariance can measure similarity between correlation of variables and correlation of classification structures. In addition, this covariance can measure the dissimilarity between variables included the dissimilarity between the classification structures for the fixed two variables. Section 3 presents the PCA based on the covariance of variables with correlation of classification structures. Section 4 describes the variable selection criterion using the fuzzy clustering result. This criterion can show how the classification structure obtained by dissimilarity of objects at each variable can match to the given classification as the external information to the data. According to the value of this criterion for each variable, we can select the significant variables. In addition, based on these selected variables, we discuss how to obtain the data in which the number of objects is larger than the number of variables which is an ordinary type of data. In order to obtain this data, we exploit the representation of interval-valued data. Based on a fuzzy clustering result from the data of the selected variables, Section 5 describes a transformation method from single-valued data to interval-valued data in order to obtain ordinary form data which is the data in which the number of objects is larger than the number of variables. Based on the ordinary form interval-valued data, we show how to obtain a fuzzy clustering result. Section 6 mentions the PCA based on covariance with weights of the fuzzy clustering result of the interval-valued data. Section 7 shows several numerical examples of the two proposed PCA for a microarray data. Finally, section 8 is the conclusion.

2 Covariance of Variables with Correlation of Classification Structures

Suppose we obtain a data as a high-dimension low-sample size data. That is, in this data the number of attributes (dimensions) is very much larger than the number of objects. We denote this situation as $p >> n$. The data matrix of variables with respect to objects as follows:

$$X = (x_{ai}), \;\; a = 1, \cdots, p, \; i = 1, \cdots, n, \;\; p >> n. \tag{1}$$

Then we define a new covariance of variables as follows [15]:

$$c_{ab} \equiv s^*_{ab} + \hat{s}^*_{ab}, \tag{2}$$

where s^*_{ab} shows correlation of degree of belongingness of variables a and b over the K clusters and $\hat{s}^*_{ab}$ shows correlation of variables a and b over n objects. That is,

$$s^*_{ab} \equiv \frac{1}{K-1}\sum_{k=1}^{K} u^*_{ak}u^*_{bk}, \;\; \hat{s}^*_{ab} \equiv \frac{1}{n-1}\sum_{i=1}^{n} x^*_{ai}x^*_{bi},$$

where

$$u^*_{ak} \equiv \frac{u_{ak} - \frac{1}{K}}{\sigma^{(u)}_a}, \; x^*_{ai} \equiv \frac{x_{ai} - \bar{x}_a}{\sigma^{(x)}_a}, \tag{3}$$

$$\sigma^{(u)}_a = \sqrt{\frac{\sum_{k=1}^{K}(u_{ak} - \frac{1}{K})^2}{K-1}}, \;\; \sigma^{(x)}_a = \sqrt{\frac{\sum_{i=1}^{n}(x_{ai} - \bar{x}_a)^2}{n-1}}, \;\; \bar{x}_a = \frac{\sum_{i=1}^{n} x_{ai}}{n}.$$

In equation (3), u_{ak} is a fuzzy grade which represents the degree of belongingness of a variable a to a cluster k, and satisfies the conditions, $u_{ak} \geq 0, \;\; \sum_{k=1}^{K} u_{ak} = 1$. u_{ak} is obtained as a fuzzy clustering result. Now we define a dissimilarity d^*_{ab} is as follows [10]:

$$\begin{aligned} d^*_{ab} &= \frac{1}{K-1}\sum_{k=1}^{K}(u^*_{ak} - u^*_{bk})^2 \frac{1}{p-1}\sum_{i=1}^{n}(x^*_{ai} - x^*_{bi})^2 \\ &= 4\{1 + s^*_{ab}\hat{s}^*_{ab} - (s^*_{ab} + \hat{s}^*_{ab})\}. \end{aligned} \tag{4}$$

That is, d^*_{ab} shows dissimilarity between variables a and b and this dissimilarity measures not only the dissimilarity between the data with respect to the variables, but also the dissimilarity between two classification structures of variables a and b. This dissimilarity is based on fuzzy self-organized dissimilarity. [10] Then, from equation (4), the correlation between variables a and b shown in equation (2) can be rewritten as follows:

$$c_{ab} \equiv s^*_{ab} + \hat{s}^*_{ab} = s^*_{ab}\hat{s}^*_{ab} - \frac{d^*_{ab}}{4} + 1. \tag{5}$$

From equations (4) and (5), we can see that this correlation can measure the similarity between the correlation of data and the correlation of classification structures for fixed variables a and b, and the dissimilarity between the variables including dissimilarity between two classification structures for the variables a and b.

Note that ordinarily PCA uses the correlation matrix with respect to variables of data, $\hat{S} = (\hat{s}^*_{ab})$, to obtain the principal components. However, in the case of the high-dimension low-sample size data, we cannot obtain any correct solution by using only $\hat{S}$. However, by adding the s^*_{ab} which is the correlation of degree of belongingness of variables a and b over the K clusters to the ordinary correlation $\hat{s}^*_{ab}$, we can use the correlation with respect to variables and obtain the solution of the PCA. This is the merit for using the proposed correlation shown in equation (2) instead of the ordinary correlation $\hat{s}^*_{ab}$.

3 PCA Based on Covariance of Variables with Correlation of Classification Structures

The α-th principal component $\boldsymbol{z}_\alpha$ of X $(p >> n)$ is defined as follows:

$$\boldsymbol{z}_\alpha = X^t \boldsymbol{l}_\alpha, \tag{6}$$

where $\boldsymbol{l}^t_\alpha = (l_{\alpha 1}, l_{\alpha 2}, \cdots, l_{\alpha p})$, and $\boldsymbol{l}_\alpha$ satisfies the condition, $\boldsymbol{l}^t_\alpha \boldsymbol{l}_\alpha = 1$. $\boldsymbol{l}_\alpha$ is obtained as the corresponding eigen-vector for the α-th largest eigen-value of $C = (c_{ab})$ shown in equation (2). The following is the algorithm of this PCA.

[Step 1] Set the obtained data shown in equation (1).

[Step 2] Set the number of clusters K. Apply the data shown in equation (1) to a fuzzy clustering method and obtain the fuzzy clustering result.

[Step 3] Using the obtained fuzzy clustering result in step 2 and the data shown in equation (1), calculate the normalized values shown in equation (3).

[Step 4] Using the obtained normalized values in step 3 and equation (4), calculate covariance shown in equation (5). Apply the obtained covariance matrix to ordinary principal component analysis and obtain the result.

4 Variable Selection Based Fuzzy Clustering

Suppose the observed data $X = (x_{ai})$ shown in equation (1) has external information for classification as data is labeled into $\hat{K}$ classes. The labeled data are shown as follows:

$$X_{\hat{k}} = (x_{ai_{\hat{k}}}), \ \ i_{\hat{k}} = 1, \cdots, n_{\hat{k}}, \ \ a = 1, \cdots, p, \ \ \hat{k} = 1, \cdots, \hat{K}, \tag{7}$$

where $\sum_{\hat{k}=1}^{\hat{K}} n_{\hat{k}} = n$. Objects in X is ordered according to the label's order. We propose a variable selection criterion to reduce the number of variables based on the external information of the classification as follows:

$$C(a) = \frac{1}{n}\left(\sum_{i_1=1}^{n_1} u_{i_1 1 a} + \cdots + \sum_{i_{\hat{K}}=1}^{n_{\hat{K}}} u_{i_{\hat{K}} \hat{K} a}\right), \ \ a = 1, \cdots, p, \tag{8}$$

where $u_{i_{\hat{k}} \hat{k} a}$ shows degree of belongingness of an object $i_{\hat{k}}$ to a class $\hat{k}$ with respect to a variable a. The object $i_{\hat{k}}$ corresponds to an object labeled to a class $\hat{k}$ which is represented as $\boldsymbol{x}_{i_{\hat{k}}} = (x_{1i_{\hat{k}}}, \cdots, x_{pi_{\hat{k}}})^t$ in equation (7). $u_{i_{\hat{k}}\hat{k}a}$ is assumed to satisfy the following conditions:

$$u_{i_{\hat{k}}\hat{k}a} \in [0,1], \ \ \forall i_{\hat{k}}, \hat{k}, a, \quad \sum_{\hat{k}=1}^{\hat{K}} u_{i_{\hat{k}}\hat{k}a} = 1, \ \ \forall i_{\hat{k}}, a. \tag{9}$$

From equation (9), the criterion shown in equation (8) can show how the obtained classification structure at each variable adjusts to the given external classification structure and $0 \leq C(a) \leq 1$. The larger value of $C(a)$ shows the greater explanatory power for the external classification information. Therefore, using a threshold for $C(a)$, we can select variables capable of explaining the external classification information of data. In order to obtain the clustering results $u_{i_{\hat{k}}\hat{k}a}$, we use a fuzzy clustering. We use the u_{ika} as a general notation. Suppose d_{ija} is (i,j)-th element of a distance matrix D_a and shows dissimilarity between objects i and j with respect to a variable a. This is defined as follows:

$$D_a = (d_{ija}), \ d_{ija} = \sqrt{(x_{ai} - x_{aj})^2}, \ \ i,j = 1, \cdots, n, \ \ a = 1, \cdots, p. \tag{10}$$

For the fuzzy clustering method in which the target data d_{ij} is dissimilarity data, the fanny method [9] is used. The objective function of this method is defined as follows:

$$J(\tilde{U}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n}\sum_{j=1}^{n} (\tilde{u}_{ik})^m (\tilde{u}_{jk})^m d_{ij} / 2 \sum_{s=1}^{n} (\tilde{u}_{sk})^m \right). \tag{11}$$

Where, $\tilde{u}_{ik}$ shows degree of belongingness of an object i to a cluster k and satisfies the conditions, $\tilde{u}_{ik} \in [0,1], \ \forall i,k, \ \ \sum_{k=1}^{K} \tilde{u}_{ik} = 1, \forall i$. m, $(1 < m < \infty)$ shows a control parameter which can control fuzziness of the belongingness. d_{ij} shows dissimilarity between objects i and j. The purpose of this method is to estimate $\tilde{U} = (\tilde{u}_{ik})$ which minimize equation (11). In equation (11), the objective function with respect to a variable a is redefined by using (10) as follows [11]:

$$J(U_a) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n}\sum_{j=1}^{n} (u_{ika})^m (u_{jka})^m d_{ija} / 2 \sum_{s=1}^{n} (u_{ska})^m \right), \ \ a = 1, \cdots, p. \tag{12}$$

Where U_a, $(a = 1, \cdots, p)$ is a matrix for a-th variable whose element u_{ika} shows degree of belongingness of an object i to a cluster k with respect to a variable a. u_{ika} can be estimated by minimizing equation (12) under the conditions, $u_{ika} \in [0,1], \ \ \forall i,k,a, \ \ \sum_{k=1}^{K} u_{ika} = 1, \ \ \forall i, a$.

5 Transformation to Interval-Valued Data

If the number of variables p is extremely large when compared with the number of objects n $(p >> n)$, then the variable selection has a problem; when the threshold value for $C(a)$ is large, loss of the data information will be large, consequently the selected variables are not sufficient to explain the data structure. Likewise, when the threshold value for $C(a)$ is small, then still we have the problem of $p > n$. In order to solve this problem, we propose a method to summarize the selected variables and transform the remained data after the variable selection to form data as $p < n$. Suppose the remained data after the variable selection shown in the previous section is as follows:

$$\tilde{X} = (\tilde{x}_{\tilde{a}i}), \ \ i = 1, \cdots, n, \ \tilde{a} = 1, \cdots, \tilde{p}, \tag{13}$$

where $\tilde{p} < p$, however it is still $\tilde{p} > n$. First, we transform the data to include the external classification information of objects. We use interval to represent each class with respect to each variable as follows:

$$Y = (y_{\tilde{a}\hat{k}}) = ([\underline{y}_{\tilde{a}\hat{k}}, \overline{y}_{\tilde{a}\hat{k}}]), \ \ \tilde{a} = 1, \cdots, \tilde{p}, \ \hat{k} = 1, \cdots, \hat{K}, \tag{14}$$

where $y_{\tilde{a}\hat{k}} = [\underline{y}_{\tilde{a}\hat{k}}, \overline{y}_{\tilde{a}\hat{k}}]$ shows the interval-valued data of the $\tilde{a}$-th variable with respect to a class $\hat{k}$ which has the minimum value $\underline{y}_{\tilde{a}\hat{k}}$ and the maximum value $\overline{y}_{\tilde{a}\hat{k}}$. From equations (7) and (13), $\underline{y}_{\tilde{a}\hat{k}}$ and $\overline{y}_{\tilde{a}\hat{k}}$ are obtained as follows:

$$\underline{y}_{\tilde{a}\hat{k}} = \min_{i_{\hat{k}}} \tilde{x}_{\tilde{a}i_{\hat{k}}}, \ \ \overline{y}_{\tilde{a}\hat{k}} = \max_{i_{\hat{k}}} \tilde{x}_{\tilde{a}i_{\hat{k}}}, \ \ \tilde{a} = 1, \cdots, \tilde{p}. \tag{15}$$

Equation (15) means that $\hat{K}$ classes over the objects which is given as external classification information are expressed by $\hat{K}$ intervals. Since the purpose of this study is identifying a subspace spanned by variables so that the subspace has strong discriminative power adjusted for the externally given classification structure of data, we assume that the given classification structure has a well separated structure. That is, we do not consider the outliers of data for classes. Although the interval representation of data is sometimes sensitive for the outliers of data, this is the reason why we can use the interval representation to the classes based on the given classification structure. In order to obtain the similarity structure of variables over the $\hat{K}$ classified objects, we classify the data shown in equation (14). The dissimilarity between $\boldsymbol{y}_{\tilde{a}} = (y_{\tilde{a}1}, \cdots, y_{\tilde{a}\hat{K}})$ and $\boldsymbol{y}_{\tilde{b}} = (y_{\tilde{b}1}, \cdots, y_{\tilde{b}\hat{K}})$ [4] is defined as

$$d_{\tilde{a}\tilde{b}} = \sum_{\hat{k}=1}^{\hat{K}} \sup\{d(x, y_{\tilde{b}\hat{k}}) | x \in y_{\tilde{a}\hat{k}}\}, \ d(x, y_{\tilde{b}\hat{k}}) = \inf\{d(x, y) | y \in y_{\tilde{b}\hat{k}}\} \tag{16}$$

and

$$d_{\tilde{b}\tilde{a}} = \sum_{\hat{k}=1}^{\hat{K}} \sup\{d(y_{\tilde{a}\hat{k}}, y) | y \in y_{\tilde{b}\hat{k}}\}, \ d(y_{\tilde{a}\hat{k}}, y) = \inf\{d(x, y) | x \in y_{\tilde{a}\hat{k}}\}. \tag{17}$$

Where, $d(x, y)$ shows distance between x and y, $\forall x \in y_{\tilde{a}\hat{k}}$, $\forall y \in y_{\tilde{b}\hat{k}}$. Therefore, $d_{\tilde{a}\tilde{b}} \neq d_{\tilde{b}\tilde{a}}$, $(\tilde{a} \neq \tilde{b})$. We use the symmetric part of the dissimilarity as

$\tilde{d}_{\tilde{a}\tilde{b}} = (d_{\tilde{a}\tilde{b}} + d_{\tilde{b}\tilde{a}})/2$. Applying this dissimilarity $\tilde{d}_{\tilde{a}\tilde{b}}$ to the fanny method shown in equation (11), we obtain a fuzzy clustering result

$$\tilde{U} = (\tilde{u}_{\tilde{a}\tilde{k}}),\ \tilde{a} = 1, \cdots, \tilde{p},\ \ \tilde{k} = 1, \cdots, \tilde{K}, \tag{18}$$

under the conditions $\tilde{u}_{\tilde{a}\tilde{k}} \in [0,1],\ \forall \tilde{a}, \tilde{k},\ \sum_{\tilde{k}=1}^{\tilde{K}} \tilde{u}_{\tilde{a}\tilde{k}} = 1, \forall \tilde{a}$, where $\tilde{K}$ is a number of categories satisfied $\tilde{K} < n$. Based on the result shown in equation (18), the data shown in equation (13) is categorized into $\tilde{K}$ categories as follows:

$$\tilde{X}_{\tilde{k}} = \{\tilde{\boldsymbol{x}}_{\tilde{a}} \mid p_{\tilde{a}\tilde{k}} = 1\},\ \ \tilde{\boldsymbol{x}}_{\tilde{a}} = (\tilde{x}_{\tilde{a}1}, \ldots, \tilde{x}_{\tilde{a}n}),\ \tilde{k} = 1, \cdots \tilde{K}, \tag{19}$$

where $p_{\tilde{a}\tilde{k}}$ satisfy

$$\tilde{u}_{\tilde{a}\tilde{k}} = \max_{1 \le \tilde{k} \le \tilde{K}} \tilde{u}_{\tilde{a}\tilde{k}} \to p_{\tilde{a}\tilde{k}} = 1,\ \ \tilde{a} = 1, \ldots, \tilde{p},$$

under the condition of $\sum_{\tilde{k}=1}^{\tilde{K}} p_{\tilde{a}\tilde{k}} = 1$. In the case that $\max_{1 \le \tilde{k} \le \tilde{K}} \tilde{u}_{\tilde{a}\tilde{k}}$ is not unique, we select the first category which appears to have the maximum degree of belongingness over the categories. We rewrite the data sets $\tilde{X}_{\tilde{k}}$ shown in equation (19) as follows:

$$\tilde{X}_{\tilde{k}} = (\tilde{x}_{\tilde{a}_{\tilde{k}}i}),\ \ i = 1, \cdots, n,\ \tilde{a}_{\tilde{k}} = 1, \cdots, \tilde{p}_{\tilde{k}},\ \ \tilde{k} = 1, \cdots, \tilde{K}, \tag{20}$$

where $\sum_{\tilde{k}=1}^{\tilde{K}} \tilde{p}_{\tilde{k}} = \tilde{p}$.

In order to create the $\tilde{K} < n$ type data, variables included to the same category are summarized for a fixed object by using an interval as follows:

$$\tilde{Y} = (\tilde{y}_{i\tilde{k}}) = ([\underline{\tilde{y}}_{i\tilde{k}}, \overline{\tilde{y}}_{i\tilde{k}}]),\ \ i = 1, \cdots, n,\ \tilde{k} = 1, \cdots, \tilde{K}, \tag{21}$$

where $\tilde{y}_{i\tilde{k}} = [\underline{\tilde{y}}_{i\tilde{k}}, \overline{\tilde{y}}_{i\tilde{k}}]$ shows the interval-valued data of the i-th object with respect to a category $\tilde{k}$ which has the minimum value $\underline{\tilde{y}}_{i\tilde{k}}$ and the maximum value $\overline{\tilde{y}}_{i\tilde{k}}$. From equation (20), $\underline{\tilde{y}}_{i\tilde{k}}$ and $\overline{\tilde{y}}_{i\tilde{k}}$ are obtained as follows:

$$\underline{\tilde{y}}_{i\tilde{k}} = \min_{\tilde{a}_{\tilde{k}}} \tilde{x}_{\tilde{a}_{\tilde{k}}i},\ \ \overline{\tilde{y}}_{i\tilde{k}} = \max_{\tilde{a}_{\tilde{k}}} \tilde{x}_{\tilde{a}_{\tilde{k}}i},\ \ i = 1, \cdots, n. \tag{22}$$

Equation (22) shows that uncertainty of variables for a category with respect to a fixed object that is represented by an interval. Since $\tilde{K} < n$ in equation (21), we can apply this data to a clustering method shown in equation (11) in order to classify the objects. First, we calculate the dissimilarity between objects $\tilde{\boldsymbol{y}}_i = (\tilde{y}_{i1}, \cdots, \tilde{y}_{i\tilde{K}})$ and $\tilde{\boldsymbol{y}}_j = (\tilde{y}_{j1}, \cdots, \tilde{y}_{j\tilde{K}})$ [4] as follows:

$$d_{ij} = \sum_{\tilde{k}=1}^{\tilde{K}} \sup\{d(x, \tilde{y}_{j\tilde{k}}) | x \in \tilde{y}_{i\tilde{k}}\},\ \ d(x, \tilde{y}_{j\tilde{k}}) = \inf\{d(x,y) | y \in \tilde{y}_{j\tilde{k}}\}, \tag{23}$$

$$d_{ji} = \sum_{\tilde{k}=1}^{\tilde{K}} \sup\{d(\tilde{y}_{i\tilde{k}}, y) | y \in \tilde{y}_{j\tilde{k}}\},\ \ d(\tilde{y}_{i\tilde{k}}, y) = \inf\{d(x,y) | x \in \tilde{y}_{i\tilde{k}}\}. \tag{24}$$

Where, $d(x,y)$ shows distance between x and y, $\forall x \in \tilde{y}_{i\tilde{k}}$, $\forall y \in \tilde{y}_{j\tilde{k}}$. From equations (23) and (24), $d_{ij} \neq d_{ji}$, $(i \neq j)$. We use the symmetric part of the dissimilarity as follows: $\tilde{\tilde{d}}_{ij} = (d_{ij} + d_{ji})/2$. Applying this dissimilarity $\tilde{\tilde{d}}_{ij}$ to the fanny method shown in equation (11), we obtain a fuzzy clustering result

$$\tilde{\tilde{U}} = (\tilde{\tilde{u}}_{ik}),\ i = 1, \cdots, n,\ \ k = 1, \cdots, K, \tag{25}$$

under the conditions,

$$\tilde{\tilde{u}}_{ik} \in [0,1],\ \forall i, k,\ \sum_{k=1}^{K} \tilde{\tilde{u}}_{ik} = 1, \forall i, \tag{26}$$

where K is a number of clusters satisfying $K < n$.

6 PCA Based on Covariance with Weights of Fuzzy Clustering Result

First, we discuss single-valued PCA which is interpreted geometrically as finding a projected space spanned by vectors that show direction of the principal components. Let L be a nonempty subset of the inner product space X. Then we define a mapping P_L from X into the subsets of L called the metric projection onto L. Then $P_L(\boldsymbol{o}_1)$ is defined as follows:

$$P_L(\boldsymbol{o}_1) = \{\boldsymbol{o}_2 \in L |\ \| \boldsymbol{o}_1 - \boldsymbol{o}_2 \| = d(\boldsymbol{o}_1, L)\},$$

where $\boldsymbol{o}_1 \in X$ and $d(\boldsymbol{o}_1, L) = \inf_{\boldsymbol{o}_2 \in L} \| \boldsymbol{o}_1 - \boldsymbol{o}_2 \|$. Let L be a convex Chebyshev set in which for each $\boldsymbol{o}_1 \in X$, there exists at least one nearest point in L. Then P_L is nonexpansive, that is,

$$\| P_L(\boldsymbol{o}_1) - P_L(\boldsymbol{o}_2) \| \leq \| \boldsymbol{o}_1 - \boldsymbol{o}_2 \|,\ \ \forall \boldsymbol{o}_1, \boldsymbol{o}_2 \in X. \tag{27}$$

The problem of the PCA is that the metric projection only satisfies equation (27) and PCA does not consider the size of values shown as follows [12], [14]:

$$C(\boldsymbol{o}_1, \boldsymbol{o}_2) = \| \boldsymbol{o}_1 - \boldsymbol{o}_2 \| - \| P_L(\boldsymbol{o}_1) - P_L(\boldsymbol{o}_2) \| .$$

Our obtained data is interval-valued data. The empirical joint density function for bivariate a and b for interval-valued data has been defined [3], [4] as follows:

$$f(\tilde{y}_k, \tilde{y}_l) = \frac{1}{n} \sum_{i=1}^{n} I_i(\tilde{y}_k, \tilde{y}_l)/||Z(i)||, \tag{28}$$

where $I_i(\tilde{y}_k, \tilde{y}_l)$ is the indicator function where each element of $(\tilde{\boldsymbol{y}}_k, \tilde{\boldsymbol{y}}_l)$ is or is not in the rectangle $Z(i) = \tilde{y}_{ik} \times \tilde{y}_{il}$ consisted of two sides which are intervals $[\underline{\tilde{y}}_{ik}, \overline{\tilde{y}}_{ik}]$ and $[\underline{\tilde{y}}_{il}, \overline{\tilde{y}}_{il}]$. $\tilde{y}_k$ and $\tilde{y}_l$ are random variables. $||Z(i)||$ is the area of this rectangle. $\tilde{\boldsymbol{y}}_k$ is k-th column vector of $\tilde{Y}$ in equation (21) and is shown as follows:

$\tilde{\boldsymbol{y}}_k = (\tilde{y}_{1k}, \cdots, \tilde{y}_{nk})^t = ([\tilde{\underline{y}}_{1k}, \overline{\tilde{y}}_{1k}], \cdots, [\tilde{\underline{y}}_{nk}, \overline{\tilde{y}}_{nk}])^t$. We extend the empirical joint density function shown in equation (28) as follows [12], [14]:

$$\tilde{f}(\tilde{y}_k, \tilde{y}_l) = \frac{1}{n} \sum_{i=1}^{n} (w_i I_i(\tilde{y}_k, \tilde{y}_l) / ||Z(i)||,$$

$$w_i = \sum_{k=1}^{K} u_{ik}^m / \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m, \quad i = 1, \cdots, n, \ \ m \in (1, \infty), \tag{29}$$

where $u_{ik}, \ i = 1, \cdots, n, \ k = 1, \cdots, K$ show the obtained degree of belongingness of the objects to the clusters when K is the selected appropriate number of clusters. Then fuzzy covariance for interval-valued data between variables k and l is derived as follows:

$$\hat{c}_{kl} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\tilde{y}_k - \bar{\tilde{y}}_k)(\tilde{y}_l - \bar{\tilde{y}}_l) \tilde{f}(\tilde{y}_k, \tilde{y}_l) d\tilde{y}_k d\tilde{y}_l, \ \ \bar{\tilde{y}}_k = \frac{1}{2n} \sum_{i=1}^{n} (\tilde{\underline{y}}_{ik} + \overline{\tilde{y}}_{ik}). \tag{30}$$

Substituting equation (29) into equation (30), and from equation (26), we have obtained the following [12], [14]:

$$\begin{aligned} \hat{c}_{kl} = (1/(4n)) \sum_{i=1}^{n} w_i (\overline{\tilde{y}}_{ik} + \tilde{\underline{y}}_{ik})(\overline{\tilde{y}}_{il} + \tilde{\underline{y}}_{il}) - (1/n) \bar{\tilde{y}}_l \sum_{i=1}^{n} (w_i (\overline{\tilde{y}}_{ik} + \tilde{\underline{y}}_{ik}))/2 \\ -(1/n) \bar{\tilde{y}}_k \sum_{i=1}^{n} (w_i (\overline{\tilde{y}}_{il} + \tilde{\underline{y}}_{il}))/2 + (1/n) \bar{\tilde{y}}_k \bar{\tilde{y}}_l. \end{aligned} \tag{31}$$

From equations (26) and (29), w_i satisfy the following condition:

$$w_i > 0, \ \ \sum_{i=1}^{n} w_i = 1. \tag{32}$$

In a hard clustering when $u_{ik} \in \{0, 1\}$, $\sum_{k=1}^{K} u_{ik} = 1$ is satisfied, the weights w_i in equation (29) is

$$w_i = 1/n, \ \ \forall i. \tag{33}$$

Since u_{ik} satisfies the conditions shown in equation (26), the weight w_i in equation (29) shows how an object is clearly classified for the obtained classification structure. If an object i is clearly classified to a cluster, then the weight w_i becomes larger, and if the classification situation with respect to an object i is an uncertainty situation, then the value of w_i becomes smaller. Therefore, it can be seen that the weights shown in equation (29) show a degree of fuzziness of the clustering with respect to each object and the proposed fuzzy covariance matrix for interval-valued data, $\hat{C} = (\hat{c}_{kl}), \ k, l = 1, \cdots, \tilde{K}$ shown in equation (31) involve a classification structure over the variables which is obtained by reflecting the dissimilarity structure of objects in a higher dimensional space shown as $\| \boldsymbol{o}_1 - \boldsymbol{o}_2 \|$ in equation (27). Therefore, based on the covariance matrix, we obtain principal components as the solution in which we can solve the problem of the ordinary PCA. The following is the algorithm for this PCA.

[Step 1] Set the obtained data shown in equation (1) to the form of data shown in equation (7).

[Step 2] Set the number of clusters K and determine the value of the control parameter m. Apply the data shown in equation (7) to a fuzzy clustering method shown in equation (12) and obtain the fuzzy clustering result $U_a,\ a = 1, \cdots, p$.

[Step 3] Using the obtained $U_a,\ a = 1, \cdots, p$ in step 2, calculate the criterion of variable selection shown in equation (8). Using the obtained values of $U(a)$, select variables which satisfy $U(a) > \varepsilon$, where ε is given.

[Step 4] Using the selected variables in step 3, formulate the data shown in equation (13). Create the interval-valued data shown in equation (14) by using the data shown in equation (13) and values shown in equation (15).

[Step 5] Calculate distance between variables shown in equations (16), (17) and obtain the symmetric part of this distance.

[Step 6] Set the number of categories $\tilde{K}$ and determine the value of the control parameter m. Apply the obtained distance in step 5 to the fuzzy clustering shown in equation (11) and obtain the fuzzy clustering result shown in equation (18).

[Step 7] Using the obtained result in step 6, formulate the data shown in equations (19) and (20). Create the interval-valued data shown in equation (21) by using the data shown in equation (20) and values shown in equation (22).

[Step 8] Calculate distance between objects shown in equations (23), (24) and obtain the symmetric part of this distance.

[Step 9] Set the number of clusters K and determine the value of the control parameter m. Apply the obtained distance in step 8 to the fuzzy clustering shown in equation (11) and obtain the fuzzy clustering result shown in equation (25).

[Step 10] Using the fuzzy clustering result in step 9, calculate the weights w_i shown in equation (29). Using the calculated weights and the data shown in equation (21), calculate the covariance shown in equation (31). Apply the obtained covariance matrix to ordinary principal component analysis and obtain the result.

7 Numerical Example

We use gene expression data for prostate cancer. [16] The data consists of 32 objects (subjects) with respect to 12626 variables (genes) shown in equation (1). As external classification information, 32 objects are labeled into two clusters of which 23 objects (microarrays) are based on mRNA extracted from microdissection of tumor tissue and 9 objects (microarrays) from normal tissue mRNA. The purpose is to identify variables (genes) and obtain the categories of variables (genes) which explain the classification structure of the two classes (a class of 23 objects with cancer and a class of 9 objects without cancer).

Using the variable selection criterion shown in equation (8), we obtained values of the criterion for each variable (gene). Figure 1 shows frequency distribution of the criterion values. Figure 2 shows variance of values of the criterion for each range. From figures 1 and 2, we can see that variance of the values which are larger than 0.8 is significantly smaller when compared with other ranges. This means that robustness for the selection of the threshold value is strong when we select the values which are larger than 0.8. Therefore, we selected variables which have more than 0.8 for the criterion. 90 variables (genes) are selected. Based on the classification of variables described in section 3, we obtained 6 categories from 90 variables. The number of categories is determined based on clarity of the clustering results.

Using the transformed 32×6 interval-valued data shown in equation (21), in order to check the classification ability, figure 3 shows the result of the proposed clustering method shown in equation (25) when $m = 2$ in equation (11). The number of clusters is determined as 2 based on a criterion shown in [12], [13]. In this figure, objects 1-23 show 23 objects (microarrays) are based on mRNA extracted from microdissection of tumor tissue and objects 24-32 show 9 objects (microarrays) from normal tissue mRNA. The value of the ordinate shows the degree of belongingness of objects to each cluster. From this figure, it can be seen that the proposed clustering method successfully classified the two classes. This result shown in figure 3 can be used for the prediction of a new object. That is, since we know the selected 90 variables are effective for the discrimination of two classes, we just need to observe the values of the 90 variables (genes). According to the obtained 6 categories of the 90 variables, we create the interval-valued data for the new object using equation (13). Adding the newly obtained interval-valued data to the original data set and applying it to the clustering method shown in section 3, we can obtain the result of fuzzy clustering shown in equation (25) for the new object. From this result, we can discriminate to which classes this object belongs. That is, we can identify the classifier as the result of fuzzy clustering shown in figure 3.

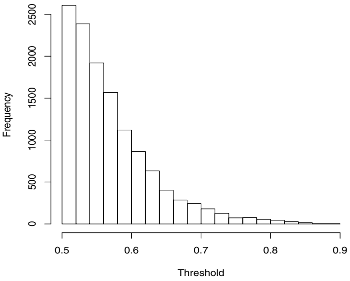


Fig. 1. Frequency Distribution of Criterion Values

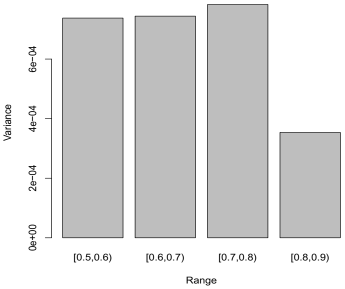


Fig. 2. Variance of Criterion Values

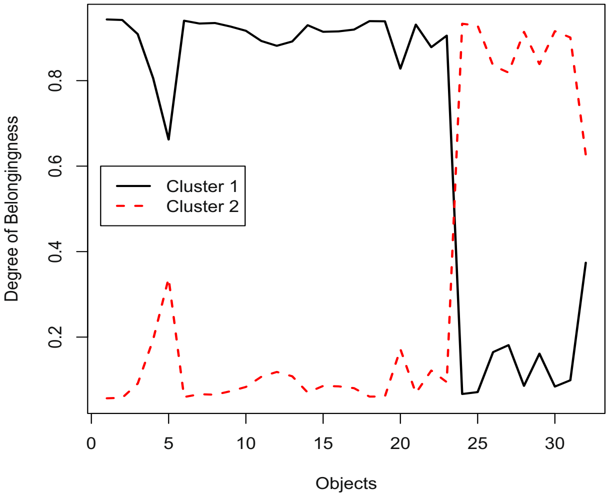


Fig. 3. Result for Proposed Clustering Method

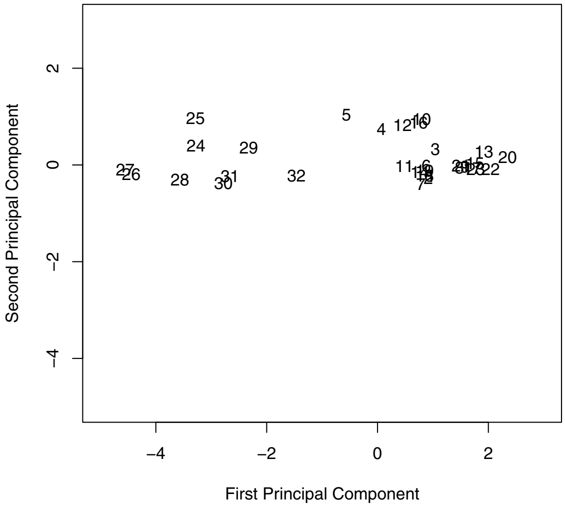


Fig. 4. Result for Proposed PCA shown in Section 6

From the obtained data shown in equation (1) and obtained weights from the result of fuzzy clustering shown in equation (29), we obtained the covariance shown in equation (31). Using this covariance, we obtain the principal components shown in figure 4. From this figure, we can see the objects classified into two clusters; 1-23 are from shavings of prostate tissue with cancer and 24-32 are from shavings of prostate tissue without cancer.

Table 1. Comparison of Cumulative Proportion

Proposed PCA	Centers Method
0.99	0.94

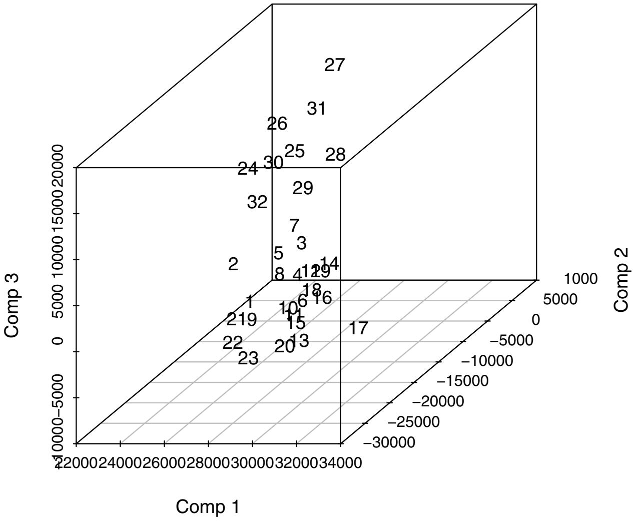


Fig. 5. Result of Proposed PCA shown in Section 3

Table 1 shows a comparison of values of cumulative proportion which is the sum of the first and the second proportions corresponding to the first and the second principal components shown in the result of figure 4 and the result of the centers method. [3], [4] This is a method of applying the data consisting of centers of intervals to the conventional PCA. This method is also identical with a method in which we use the conventional empirical joint density function shown in equation (28) and derive the covariance and then apply the obtained covariance into the conventional PCA. From equation (33), this method is the same as a case in which we use a hard clustering in a high dimensional space in our proposed PCA. From equations (32) and (33), for the fair comparison of fuzzy and hard clustering, we multiplied n to equation (31). From the result shown in table 1, we can see that the proposed PCA could obtain a better result.

Figure 5 shows the result of PCA shown in equation (6). From the result shown in figure 5, we can see that the objects are successfully classified into the two given groups. That is, in this figure, 24 - 32 show objects without cancer and the other numbers show objects with cancer. Note that in this case, we do not use any external information of the classes in the data, however, we can obtain the adaptable clustering result by using the proposed method shown in section

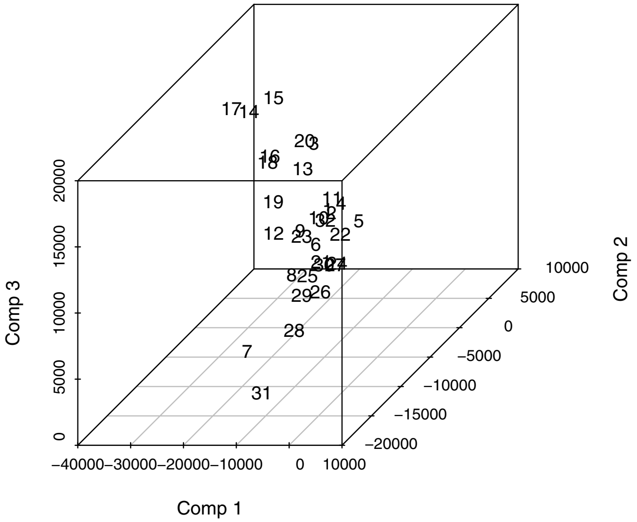


Fig. 6. Result of Ordinary PCA

3. Figure 6 shows a result of ordinary PCA. From this figure, we cannot see any clear classification of the two given groups.

8 Conclusions

This paper presents principal component analyses based on covariances harnessing fuzzy clustering for high dimension low sample-size data. Numerical examples show a better performance by applying microarray data which is a typical high dimension low sample-size data to the proposed methods.

References

1. Ahn, J., Marron, J.S., Muller, K.M., Chi, Y.-Y.: The High-Dimension, Low-Sample-Size Geometric Representation Holds under Mild Conditions. Biometrika 94(3), 760–766 (2007)
2. Baik, J., Arous, G.B., Peche, S.: Transition of the Largest Eigenvalue for Nonnull Complex Sample Covariance Matrices. The Annals of Probability 33(5), 1643–1697 (2005)
3. Billard, L., Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley (2007)
4. Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data. Springer (2000)
5. Friedman, J.H., Meulman, J.J.: Clustering Objects on Subsets of Attributes. Journal of the Royal Statistical Society: Series B 66(4), 815–849 (2004)
6. Hall, P., Marron, J.S., Neeman, A.: Geometric Representation of High Dimension Low Sample Size Data. Journal of Royal Statistical Society 67(pt. 3), 427–444 (2005)

7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer (2009)
8. Johnstone, I.M.: On the Distribution of the Largest Eigenvalue in Principal Components Analysis. The Annals of Statistics 29(2), 295–327 (2001)
9. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. John Wiley & Sons (1990)
10. Sato-Ilic, M., Kuwata, T.: On Fuzzy Clustering based Self-Organized Methods. In: FUZZ-IEEE 2005, pp. 973–978 (2005)
11. Sato-Ilic, M.: Fuzzy Variable Selection with Degree of Classification based on Dissimilarity between Distributions of Variables. International Journal of Intelligent Technology and Applied Statistics 1(2), 1–18 (2008)
12. Sato-Ilic, M.: A Cluster-Target Similarity Based Principal Component Analysis for Interval-Valued Data. In: 19th International Conference on Computational Statistics, pp. 1605–1612. Physica-Verlag (2010)
13. Sato-Ilic, M.: Clustering High Dimension Low Sample-Size Data with Fuzzy Cluster-based Principal Component Analysis. In: The 58th ISI World Statistics Congress, cps024-5 (2011)
14. Sato-Ilic, M.: Symbolic Clustering with Interval-Valued Data. Procedia Computer Sciences 6, 358–363 (2011)
15. Sato-Ilic, M.: On Fuzzy Clustering based Correlation. Procedia Computer Sciences 12, 230–235 (2012)
16. Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A., Frierson Jr., H.F., Hampton, G.M.: Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer. Cancer Research 61, 5974–5978 (2001)

Random Matrix Theory and Covariance Matrix Filtering for Cancer Gene Expression

Leif E. Peterson[1] and Charles E. Ford[2]

[1] Center for Biostatistics, TMHRI
6565 Fannin, Suite MGJ6-031, Houston, Texas 77030 USA
lepeterson@tmhs.org
[2] Division of Biostatistics, University of Texas - School of Public Health
1200 Hermann Pressler Street, Houston, Texas 77225 USA
charles.e.ford@uth.tmc.edu

Abstract. We investigated random matrix theory (RMT) and covariance matrix filtering with shrinkage techniques to characterize eigendecomposition of a 190 × 190 covariance matrix based on 750 genes and 18 tumor classes. Principal component subtraction using the first PC resulted in the most favorable outcome concerning eigenvector participation ratios, class-specific influence scores, and unsupervised clustering of arrays. By fitting the Marčenko-Pastur density function, we determined that 86.8% of the covariance matrix eigenvalues were below the threshold value of $\lambda^+ = 0.5025$, suggesting that they reside in the noise region. Removal of noise eigenvector effects in the data were not as informative as removal of only the first eigenvector, however, there were interesting properties observed among the 25 non-zero eigenvalues after noise removal – mostly that they were lower than the first 25 eigenvalues of the remaining types of covariance matrices.

1 Introduction

Random matrix theory (RMT) initially became popular in nuclear physics for describing resonance fluctuations of compound nuclei [1]. At present, random matrices are employed in many fields ranging from probability theory to complexity theory [2–5]. In the field of statistics, a commonly used random matrix is the white Wishart matrix, $W_p(n, \boldsymbol{\Sigma}) = \mathbf{X}\mathbf{X}^T$, which is a square symmetric $p \times p$ matrix, and has n degrees of freedom based on the number of rows of $\mathbf{X}$ whose columns are constructed from uncorrelated and i.i.d. random variates. Examples of white Wishart matrices are the sample covariance matrix $\mathbf{C}$ or correlation matrix $\mathbf{R}$ for any $n \times p$ random data matrix $\mathbf{X}$. By definition, it is known that under zero correlation, $\mathbf{C} = \mathbf{R} = \mathbf{I}$, the determinant and eigenvalues of $\mathbf{I}$ are known to be unity. However, when the columns of $\mathbf{X}$ are random and assumed to be uncorrelated, both $\mathbf{C}$ and $\mathbf{R}$ only approach $\mathbf{I}$ as $n \to \infty$, due to the observed spread in the empirical eigenvalue distribution. For any given dataset, it is assumed that the sample covariance matrix $\mathbf{C}$ accurately represents the population covariance matrix $\boldsymbol{\Sigma}$, however, as $p \to n$ or if $p > n$, the eigenvalues become unreliable and can also take on a value of zero, resulting in lack of

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 173–184, 2013.

positive definiteness. With high-dimensional datasets becoming more popular in genomic sciences and exploratory data mining, there is greater potential for the number of dimensions to approach the sample size ($p \to n$), leading to biased eigenvalues of $\mathbf{C}$ and $\mathbf{R}$. Certainly, there will be $p-n$ zero eigenvalues whenever $p > n$ and one zero eigenvalue whenever $p = n$.

This investigation explores RMT techniques to attempt to correct for eigenvalue problems associated with a large gene expression dataset for 18 types of tumors with 190 arrays and 750 genes (mRNAs). Firstly, we address the removal of the first principal component (PC) of $\mathbf{C}$ from $\mathbf{X}$ and then redetermination of $\mathbf{C}$ in an attempt to remove widespread correlation observed in $\mathbf{X}$, which causes the greatest eigenvalue to be biased toward exceedingly high values. We then apply the Marčenko-Pastur law to identify the noise region of the empirical eigenvalue distribution, and then remove the effects of noise eigenvectors from $\mathbf{X}$, followed by redetermination of $\mathbf{C}$. This is followed by joint removal of both the first PC and noise eigenvectors from $\mathbf{X}$ and redetermination of $\mathbf{C}$. The last approach involves covariance matrix filtering via shrinkage, in which we apply several shrinkage techniques to $\mathbf{C}$ in order to directly estimate lesser-biased shrunken variants of $\mathbf{C}$.

2 Methods

2.1 Dataset

We applied RMT to a 190×190 (array-by-array) covariance matrix for 190 DNA microarrays and 750 genes identified using F-tests in the GCM expression data set for 18 classes of tumors [6]. The 18 tumor classes are BR-breast, PR-prostate, LC-lung cancer, CO-colorectal, LYF-folicular lymphoma, LYLB-large B-cell lymphoma, ME-melanoma, BL-bladder, UT-uterine, ALLB-leukemia, ALLT-leukemia, AML-acute myelocytic leukemia, RE-renal, PAN-pancreatic, OV-ovarian, MES-mesothelioma, GLI-CNS-glioblastoma, MED-CNS-medulloblastoma. The 750 genes were selected from 16,064 informative genes of the Affymetrix Hu6800 and Hu35KsubA chips identified using multiclass F-tests. Altogether, the final gene expression dataset resulted in a 750×190 $\mathbf{X}$ matrix for which $n = 750$ and $p = 190$; thus, we set up the 190 arrays as columns of $\mathbf{X}$ for covariance and correlation. Additionally, because this paper focused only on covariance analyses and not supervised classification, we did not perform further filtering to identify genes that were the best class predictors. Rather, the covariance analyses performed always included the full dataset consisting of 750 genes and 190 arrays.

Techniques used in this investigation for determining the empirical eigenvalue distribution (e.e.d.) of $\mathbf{C}$ are described in detail in [7], so only brief summaries are provided in the following paragraphs. The first type of covariance matrix we estimated was based on permuting columns of $\mathbf{X}$ to yield the the covariance matrix $\mathbf{C}_{PERM}$. When columns of $\mathbf{X}$ are permuted, their variances remain intact; however, the off-diagonals representing covariance in $\mathbf{C}$ tends to zero. When no corrections were made to covariance, we called the result $\mathbf{C}_{RAW}$. For component subtraction, gene expression $\mathbf{x}_j$ for the jth array was regressed on scores for the first principal component $\mathbf{f}_1$, where each $f_{i1} = \sum_k e_{k1} x_{ik}$ and the data vector $\mathbf{x}_j$ was replaced with the regression residuals to generate the fully replaced data matrix called $\mathbf{X}_{RES}$, which resulted in the covariance matrix $\mathbf{C}_{RES}$. The Marčenko-Pastur (MP) law [8] states that for i.i.d. columns in $\mathbf{X}$ and $(n, p \to \infty, \gamma = p/n)$, the minimum and maximum eigenvalues of $W_p(n, \boldsymbol{\Sigma})$ almost surely converge to $\lambda^- = \sigma^2(1-\sqrt{\gamma})^2$ and $\lambda^+ = \sigma^2(1+\sqrt{\gamma})^2$, respectively. The e.e.d. for $W_p(n, \boldsymbol{\Sigma})$ based on $\mathbf{X}$ with i.i.d. elements is given by

$$f(\lambda) = \max\left(0, 1 - \frac{1}{\gamma}\right)\delta(\lambda) + \frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{2\pi\gamma\lambda\sigma^2} I(\lambda^- \leq \lambda \leq \lambda^+), \qquad (1)$$

where the $\max()\delta(\lambda)$ term represents the density at $\lambda = 0$ for the $p - n = p(1 - 1/\gamma)$ zero eigenvalues when $p > n$, i.e., $\gamma > 1$, and the $I()$ represents the density when λ is between λ^- and λ^+. Particle swarm optimization was used to obtain the best fit of $f(\lambda)$ to the observed values of λ, with fitted results being estimates of γ, σ, λ^-, and λ^+. Multivariate regression was used to perform the regression analysis $\mathbf{X} = \mathbf{F}\boldsymbol{\beta}$, where $\mathbf{F}$ is the matrix of PC scores for the noise eigenvectors whose eigenvalues were less than λ^+. The regression residuals were used to replace the original gene expression values in $\mathbf{X}$. Using the new $\mathbf{X}$ matrix with noise removed, we calculated a new covariance matrix called $\mathbf{C}_{MP}$. The joint correction for widespread correlation and noise was accomplished by augmenting $\mathbf{F}$ with $\mathbf{f}_1$ and regressing $\mathbf{X}$ on the augmented matrix to yield residuals that were used to replaced the data in $\mathbf{X}$. Covariance determination of the newly defined $\mathbf{X}$ matrix resulted in the covariance matrix $\mathbf{C}_{RESMP}$.

We employed three shrinkage methods to $\mathbf{C}$ that were introduced by Daniels-Kass (DK) [9], Ledoit-Wolf (LW) [10], and Schäfer-Strimmer (SS) [11]. The DK shrinkage method shrinks eigenvalues of $\mathbf{C}$ that are assumed to be log-normally distributed and forms a maximum likelihood estimator based on the log-normal priors, where $\log(\lambda_j) \sim \mathcal{N}(\log(\lambda), \tau^2)$ and $\tau^2 = \sum_j (\log(\lambda_j) - \langle\log(\lambda)\rangle)^2/(p+4) - 2/n$. New eigenvalues are determined as $\lambda_j = \exp\{(2/n)/(2/n + \tau^2)\langle\log(\lambda)\rangle + \tau^2/(2/n + \tau^2)\lambda_j\}$. Any zero eigenvalues of $\mathbf{C}$ were replaced with the mean value of λ_j based on the number of zero eigenvalues in the original $\mathbf{C}$ matrix. DK shrinkage resulted in the covariance matrix $\mathbf{C}_{DK}$. LW shrinkage is based on $\mathbf{C}_{LW} = \delta\mathbf{C}^* + (1-\delta)\mathbf{C}$, where $\mathbf{C}^*$ is a highly structured estimator of $\mathbf{C}$, and δ is the shrinkage intensity in the range [0,1] which is the weight applied to the structured estimator. LW shrinkage resulted in the covariance matrix $\mathbf{C}_{LW}$. The SS shrinkage technique to obtain $\mathbf{C}_{SS}$ replaces diagonal elements of $\mathbf{R}$ with ones, and off-diagonals r_{jk} with $r_{jk}\min(1, \max(0, 1-\lambda^*))$, where the shrinkage intensity is $\lambda^* = \sum_{j \neq k} \mathrm{var}(r_{jk}) / \sum_{j \neq k} r_{jk}^2$. Anytime covariance was needed from correlation, we used the relationship $c_{jk} = r_{jk}\sigma_j\sigma_k$, whereas correlation was obtained from covariance in the form $r_{jk} = c_{jk}/\sqrt{c_{jj}c_{kk}}$. For each covariance matrix described above, we calculated the inverse participation ratio [12] in the form

$$\mathrm{IPR}_j = \sum_{k=1} |e_{jk}|^4, \qquad (2)$$

where e_{jk} is the eigenvector element of the jth eigenvector. Large values of IPR suggest that only several microarrays contribute to the eigenvector, whereas small values indicate that the microarrays contribute equally. A single XY scatter plot was constructed with IPR_j on the y-axis and all eigenvalues of $\mathbf{C}$ on the x-axis. We also determined the average class-specific influence on the eigenvector $\mathbf{e}_1$ associated with λ_1 by initially setting an indicator to denote the class ω of each array using

$$\varDelta_{l\omega} = \begin{cases} 1, & \text{if} \quad \mathbf{x}_l \in \omega \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and looped over all arrays to derive the influence on $\mathbf{e}_1$,

$$\mathrm{IPR}_{i\omega} = \frac{1}{n_\omega} \sum_{l=1}^{n} \varDelta_{l\omega} |e_{il}|^2 . \tag{4}$$

Plots of $\mathrm{IPR}_{i\omega}$ were constructed for all classes of arrays and $\mathbf{e}_1$ for each of the covariance matrices described above.

3 Results

Figure 1 shows the empirical distribution of covariance $\mathbf{C}$ (top) and correlation $\mathbf{R}$ (bottom) matrix elements based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage. For $\mathbf{C}_{RAW}$, all of the covariance values were positive, suggesting that there exists a *Perron-Frobenius eigenvalue* λ_1 which is strictly larger than the other eigenvalues, as well as the existence of an eigenvector whose elements are all positive. In fact, the values of λ_1, λ_2, and λ_3 with variance explained for $\mathbf{C}_{RAW}$ were 121.2(0.64), 14.3(0.08), and 6.5 (0.03) – revealing a strictly larger eigenvalue. In addition, a check of the eigenvector $\mathbf{e}_1$ for λ_1 indicated all elements had the same sign. It is apparent that correlation values have a wider range than covariance, and there is more overlap among the correlation values, especially between the component subtraction and shrinkage results. The first and foremost observation is that removal of the effect of the first PC from the data results in covariance values (RES) that take on positive and negative values. The MP-based correction caused the distribution to shift slightly rightward into positive territory. Overall, for covariance, there was much less overlap between results of the component subtraction methods vs. shrinkage results. The shrinkage methods did not seem to have a strong effect on eigenvalues for $\mathbf{C}$ or $\mathbf{R}$.

Figure 2 shows the eigenvalues of the covariance $\mathbf{C}$ (top) and correlation $\mathbf{R}$ (bottom) matrices based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise

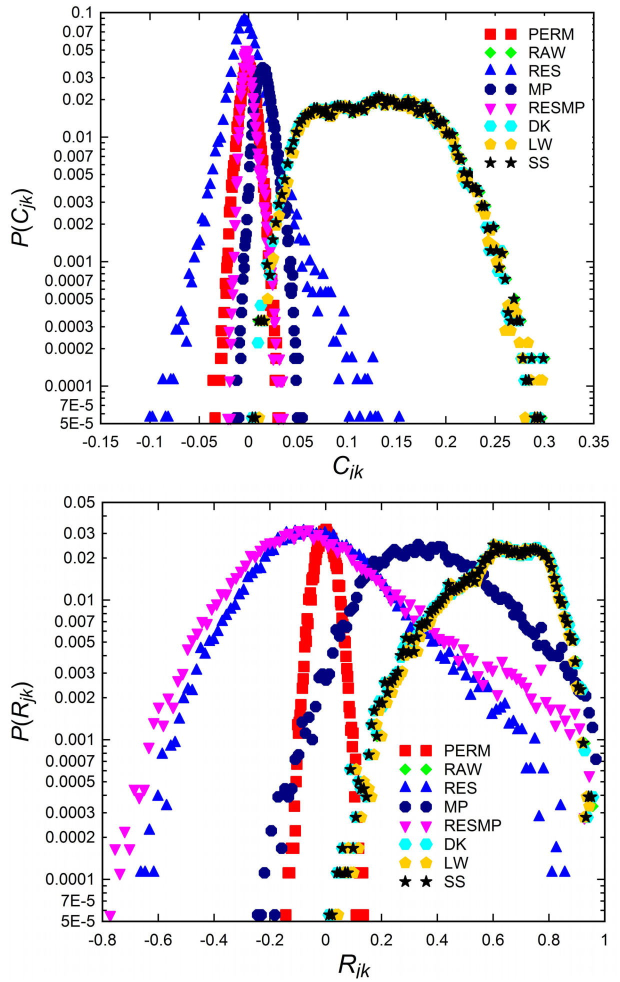


Fig. 1. Empirical distribution of elements of 190 × 190 covariance matrices **C** (top) and correlation matrices **R** (bottom) based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage.

PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage. The dynamic range and overlap of covariance-based eigenvalues is much lower than that for the correlation-based eigenvalues. There were 165 eigenvalues (86.8%=165/190) below the maximum noise threshold of $\lambda^+ = 0.5025$, which was determined by fitting the MP law to the eigenvalues of **R**. A reference line for $\lambda^+ = 0.5025$ can be seen in the bottom panel of Figure 2. After MP-based noise removal in **X** using PC scores for the 165 noise eigenvectors, there was by definition only 25 non-zero eigenvalues extracted from the newly generated **C** matrix based on the denoised **X**.

For covariance, the MP-based and RESMP-based eigenvalues are distinctly different from the bulk of eigenvalues. A major issue with eigenvalues based on correlation is that the eigenvalues for correlation of the permutation matrix (PERM) are not far-removed and cut through the remaining eigenvalues for various component subtraction and shrinkage. For covariance-based eigenvalues, however, eigenvalues for the random permutation matrix are mostly always unique and far-removed with little overlap with eigenvalues for the component subtraction and shrinkage techniques. Interestingly, the MP-based noise reduction in **X** followed by generation of a new **C** matrix resulted in eigenvalues that were more straight on a scree plot. One advantage for using **C** when compared with **R** is that the 25 non-zero eigenvalues of $\mathbf{C}_{MP}$ were mostly lower than the first 25 eigenvalues of the other **C** matrices.

Figure 3 reflects the empirical eigenvalue distribution of **R** and the fitted results based on the MP distribution. The upper limit of noise for eigenvalues was $\lambda^+ = 0.5025$, and therefore, all eigenvalues below 0.5 were assumed to be noise when performing the multivariate regression to generate residuals and then $\mathbf{C}_{RES}$. Figure 4 shows results for IPR_j for eigenvectors as a function of eigenvector of the covariance matrix **C** based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage. Larger values of IPR denote that a few arrays dominate the eigenvector while low values indicate equal contribution to the eigenvector elements. IPRs for eigenvectors related to the MP- and RESMP-based covariance matrices seem to occupy the distribution at values greater than one, with random permutation results near unity, and the remaining eigenvalues in the noise region. A majority of IPRs for RES lie near the IPRs for the random covariance (PERM), and their small magnitude in IPR indicates that the arrays contribute nearly equally to the various eigenvectors.

Figure 5 shows class-specific $\mathrm{IPR}_{i\omega}$ for $\mathbf{e}_1$ of the covariance matrix **C** based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage. IPRs for **C** of the random permutation matrix of **X** were jumpy, while those for **C** of the uncorrected data matrix (RAW) were quite similar – which implies that the majority of arrays in all classes influenced $\mathbf{e}_1$ similarly. By far, the most impressive results were obtained after removing the effect of the first PC on the data matrix **X**, which is shown in the panel for RES. Results for MP were less varied when compared with RESMP which removes effects of the first PC and noise from **X** before determining **C**. Shrinkage methods resulted in $\mathrm{IPR}_{i\omega}$ values which were similar to those for RAW.

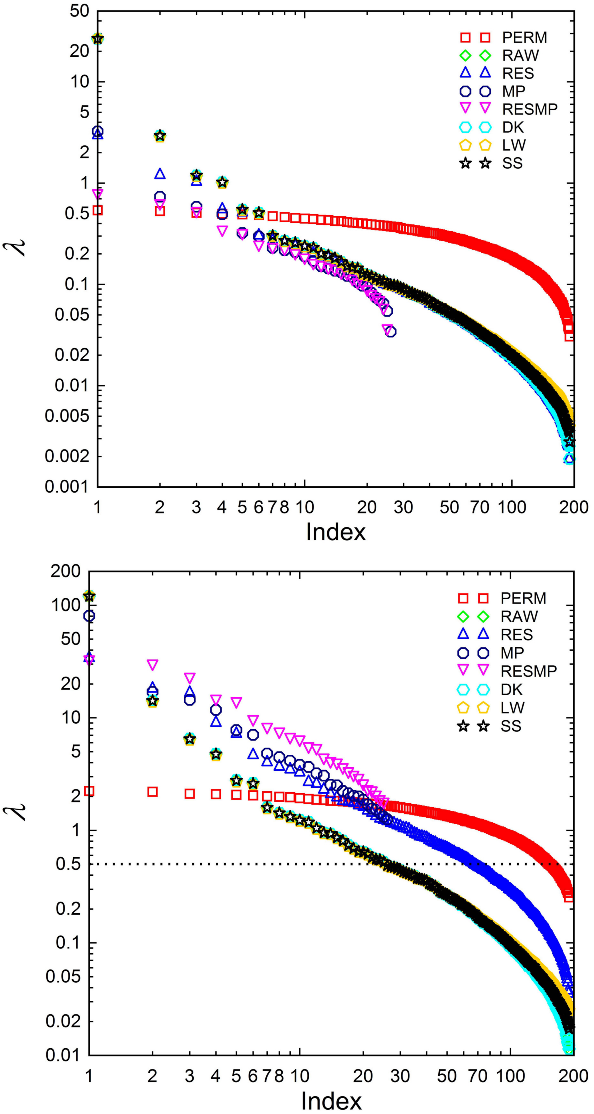


Fig. 2. Eigenvalues of the 190 × 190 covariance **C** (top) and correlation **R** (bottom) matrices based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage. A reference line for $\lambda^+ = 0.5025$ for the MP distribution can be seen in the bottom panel.

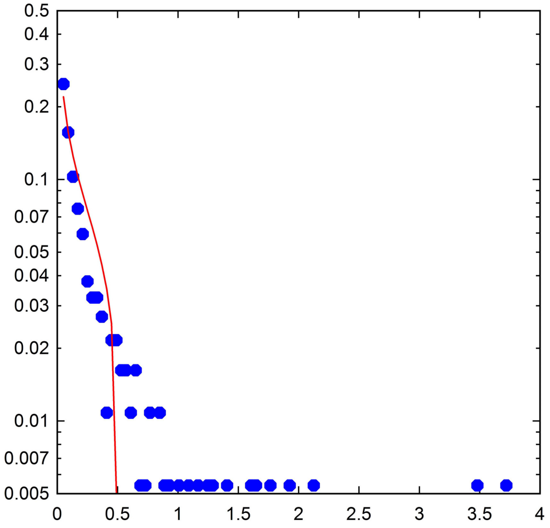


Fig. 3. Empirical eigenvalue distribution for 190 × 190 **R** showing the fitted line based on the MP distribution. Upper limit of noise threshold for eigenvalues from fitted MP distribution was $\lambda^+ = 0.5025$.

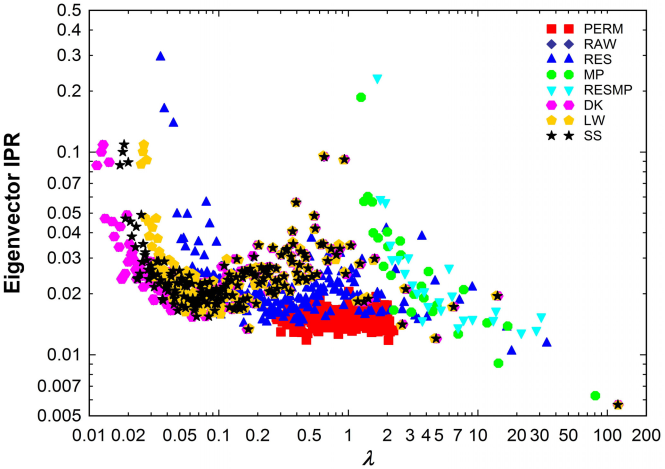


Fig. 4. Inverse participation ratios IPR_j for eigenvectors as a function of eigenvalue of the 190 × 190 covariance matrix **C** based on random permutation (PERM) data matrix, no correction (RAW), data residuals after regression on first PC (RES), data residuals after regression on noise PCs (MP), data residuals for both the first PC and noise PCs (RESMP), Daniels-Kass (DK) shrinkage (DK), Ledoit-Wolf (LW) shrinkage, and Schäfer-Strimmer (SS) shrinkage.

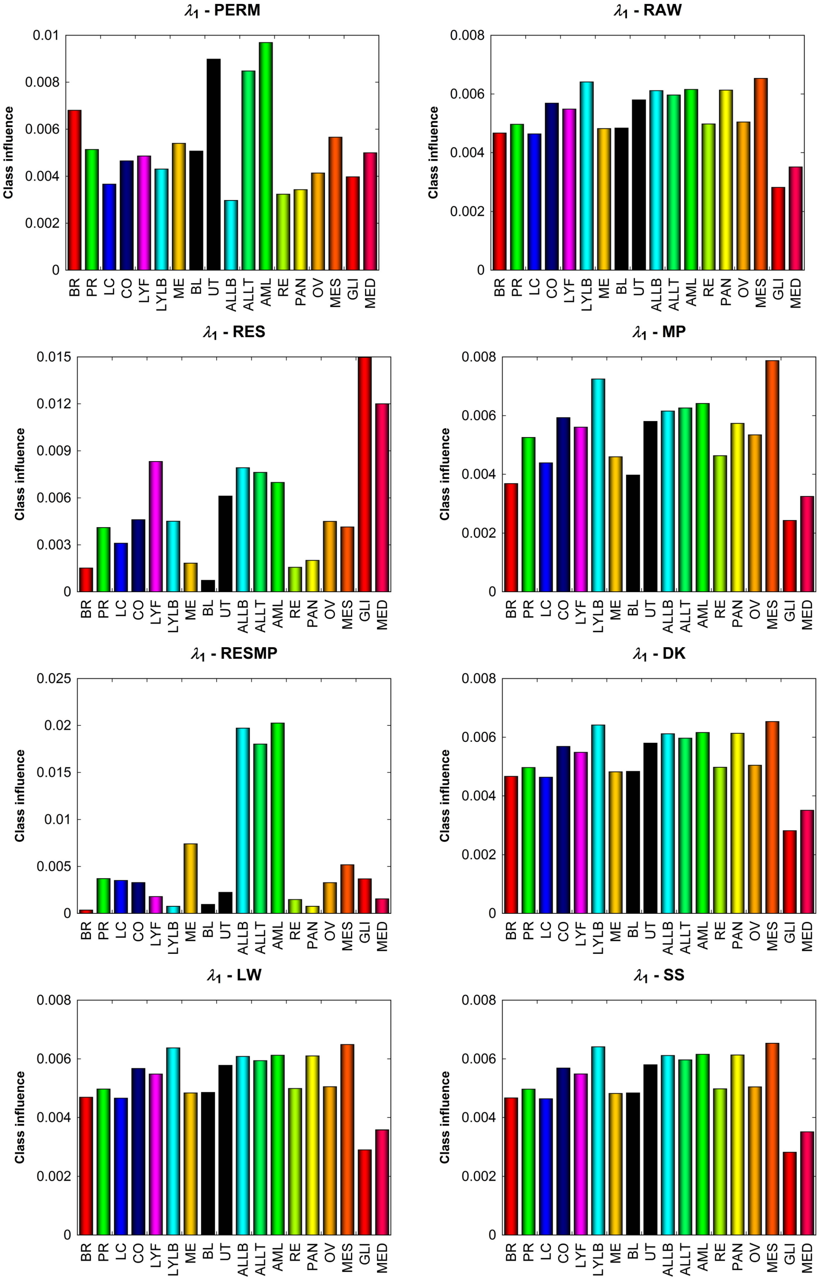


Fig. 5. Tumor class-specific influence scores $\mathrm{IPR}_{i\omega}$ for the eigenvector $\mathbf{e}_1$ of the various 190×190 covariance matrices

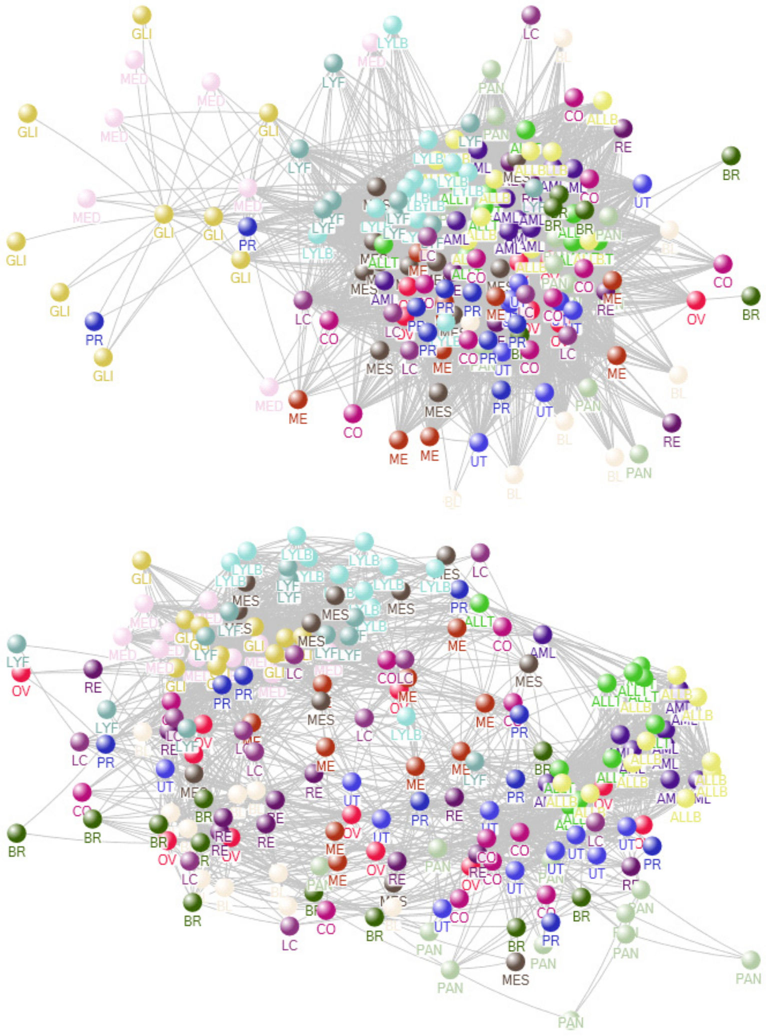


Fig. 6. Force plots for the RAW (top panel) and RES (bottom panel) 190 × 190 covariance matrix elements. Between-array covariance is more heterogeneous for RAW when compared with RES.

Figure 6 shows that, on the top panel, the RAW covariance between the 190 arrays is more heterogeneous when compared with between-array covariance based on RES (bottom panel). In the bottom panel for RES, many of the arrays in the same class are proximal with one another, especially for GLI, MED, LYYF, LYLB, ALLB, ALLT, UT, and finally ME, and BR which are slightly scattered. Almost the same relationships exist between arrays in various classes in the top panel (RAW), however, there are greater scale changes.

4 Discussion

The results obtained suggest that covariance matrices derived after component subtraction (RES, MP, and RESMP) reflect a greater change when compared with shrinkage results. Removal of the effects of the first PC caused a slight increase in skewness of the positive values of C_{jk} in the RES covariance matrix. Elements of the correlation matrices, R_{jk}, presented too much overlap between the various type of correlation matrices, such that there is less uniqueness. Hence, correlation is likely not as beneficial as covariance matrix filtering prior to post hoc discrimination analysis. The Marčenko-Pastur law was instrumental for identifying an upper limit ($\lambda^{+} = 0.5025$) of the noise threshold for eigenvalues of $\mathbf{R}$ matrix, for which there were 165 eigenvalues below this threshold.

The smaller scale of the 25 non-zero eigenvalues from $\mathbf{C}_{MP}$ have the advantage of representing lower variance, which represent less risk. Maximum ROI from eigenanalysis is realized when truly non-noise eigenvectors with near-zero eigenvalues can be found, because the dimensions that load on such eigenvectors will typically be orthogonal. The other half of the payoff is that the mean effects of the considered dimensions are high. Thus, the optimal eigenvector is comprised of uncorrelated dimensions with large mean effects – essentially resulting in a large payoff with little variance (zero eigenvalue). The most important criterion for identifying such eigenvectors is that their eigenvalues will likely be very close to the Marčenko-Pastur value of λ^{+}, and confirming their merit is a challenging task, since their existence may not be true. Inverse participation ratios (IPRs) allowed us to observe the contribution of arrays to all of the eigenvectors that were determined. Component subtraction using the RES approach resulted in a majority of IPRs that were proximal to IPRs of the PERM covariance. Thus, some of the better candidate eigenvectors have λ near 0.5 (close to λ^{+}) and have low values of IPR below 0.02. IPR values in this study that were above 0.04 reflect participation of only a few dimensions (arrays) and are therefore of less interest. Recall, since eigenvalues below 0.5 fall in the noise range, IPR results below this value are also not of great interest.

The assessment of class-specific influence scores on eigenvetor $\mathbf{e}_1$ associated with the greatest eigenvalue λ_1 also reflects that RES causes more variation among the classes, which should be more informative for discrimination. The formulation for class-specific influence scores summed array-specific eigenvector elements within each class, and we therefore did not assess the individual influence of arrays.

We applied RMT and covariance matrix filtering via component subtraction and shrinkage. All of the RAW covariance matrix elements were observed to be positive. We have shown that the e.e.d. of eight types of covariance and correlation matrices have widely varying characteristics, and primarily that 86.8% of the eigenvalues fall in the noise region. We have found that removing widespread correlation among the arrays results in better participation ratios, class-specific influence scores, and between-array clustering based on force plot construction. Our current research can address several directions, ranging from theoretical studies of RMT and covariance filtering on gene expression, to applied investigations involving discrimination of unknown test arrays. Finally, we think that our present work will allow us to appropriately denoise high-dimension gene expression datasets using linear and non-linear manifold learning, prior to root-MUSIC superresolution to exploit noise.

5 Conclusions

We investigated RMT and covariance matrix filtering with shrinkage techniques to characterize the eigenspace of a 190 × 190 covariance matrix based on 750 genes and 18 tumor classes. Principal component subtraction using the first PC resulted in the most favorable outcome concerning eigenvector participation ratios, class-specific influence scores, and unsupervised clustering of arrays. By fitting the Marčenko-Pastur density function, we determined that 86.8% of the covariance matrix eigenvalues were below the threshold value of $\lambda^+ = 0.5025$, suggesting that they reside in the noise region. Removal of noise eigenvector effects in the data were not as informative as removal of only the first eigenvector, however, there were interesting properties observed among the 25 non-zero eigenvalues after noise removal – mostly that they were lower than the first 25 eigenvalues of the remaining types of covariance matrices.

References

1. Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Annals of Mathematics 62(3), 548–564 (1955)
2. Wishart, J.: Generalized product moment distributions in samples. Biometrika 1(2), 32–52 (1928)
3. James, A.T.: Normal multivariate analysis and the orthogonal group. Annals Math. Stat. 1, 40–75 (1954)
4. Pastur, L.A.: On the spectrum of random matrices. Teor. Mat. Fiz. 10, 102–111 (1973)
5. Khorunzhy, A.: Sparse random matrices: spectral edge and statistics of rooted trees. Adv. Appl. Prob. 33, 1–18 (2001)
6. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R.: Multiclass cancer diagnosis using tumor gene expression signatures. PNAS 98(26), 15149–15154 (2001)
7. Peterson, L.E.: Classification Analysis of DNA Microarrays. John Wiley, New York (2013)
8. Marčenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Mat. Sb. (N.S.) 72(114), 4, 507–536 (1967)
9. Daniels, M.J., Kass, R.E.: Shrinkage estimators for covariance matrices. Biometrics 57(4), 1173–1184 (2001)
10. Ledoit, O., Wolf, M.: Honey, I shrunk the sample covariance matrix. J. Portfolio Management 30(4), 110–119 (2004)
11. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Applic. Genet. Molec. Biol. 4(1), 1–30 (2005)
12. Biely, C., Thurner, S.: Random matrix ensembles of time-lagged correlation matrices: derivation of eigenvalue spectra and analysis of financial time-series. Quant. Finance 8, 705–722 (2008)

Author Index

Ambrogi, Federico 117

Bassani, Niccolò 117
Biganzoli, Elia 117

Cantoni, Virginio 57
Chang, Yu-Bing 48
Chen, Peikai 132
Cucinotta, Francis A. 1, 12, 26

Fan, Yubo 132
Ferone, Alessio 57
Fiannaca, Antonino 105
Ford, Charles E. 173

Garibaldi, Jonathan M. 147
Gateno, Jaime 48

Haspel, Nurit 70
Hu, Shaowen 26
Hung, Y.S. 132

Kuo, Tai-Hong 48

Lai, Daphne Teck Ching 147
La Rosa, Massimo 105
Lau, Ching C. 132

Mahmoud, Hassan 37
Man, Tsz-kwong 132
Masulli, Francesco 37
Miyanishi, Kazunori 81

Ohkawa, Takenao 81
Ozbudak, Ozlem 57

Peterson, Leif E. 173
Petrosino, Alfredo 57
Plante, Ianik 12
Ponomarev, Artem L. 1

Rizzo, Riccardo 105
Rovetta, Stefano 37

Sato-Ilic, Mika 158
Simovici, Dan 70
Smith, Jennifer A. 95

Urso, Alfonso 105

Vetro, Rosanne 70

Wong, Stephen T.-C. 132

Xia, James J. 48
Xiong, Zixiang 48

Yuan, Peng 48

Zhou, Xiaobo 48