

Schedule

Thursday May 20

8:00-9:00	Breakfast	Village 1 Dining Room
9:00-12:30	MITACS Step Workshops	Davis Centre
12:30-13:30	Lunch	Village 1 Dining Room
13:30-16:30	Tutorials	Davis Centre
17:30	GSEF Ice Breaker	Grad House

Friday May 21

8:00-8:50	Breakfast	Village 1 Dining Room
8:50-9:00	Conference Opening	Davis Centre 1302
9:00-10:00	Keynote Speaker - Ming Li	Davis Centre 1302
10:00-10:15	Coffee Break	Davis Centre 1301
10:15-11:55	Morning Paper Presentations	Davis Centre 1302/1304
11:55-13:30	Lunch	Village 1 Dining Room
13:00-13:30	“What to expect from PhD until Assistant Professor” (informal) - Nobuhiko Hata	Davis Centre 1302
13:30-14:30	Keynote Speaker - Nobuhiko Hata	Davis Centre 1302
14:30-14:45	Coffee Break	Davis Centre 1301
14:45-16:25	Afternoon Paper Presentations	Davis Centre 1302/1304
16:25-18:00	Poster Session	Davis Centre 1302
18:00-20:30	Dinner	Concordia Club
Only transportation is provided, dinner not included		
Bus leaves from in front of DC at 18:00		

Saturday May 22

8:00-9:00	Breakfast	Village 1 Dining Room
9:00-10:00	Keynote Speaker - S. Cenk Sahinalp	Davis Centre 1302
10:00-10:45	Coffee Break / Poster Session	Davis Centre 1301
10:45-11:15	Business Meeting (open to everyone)	Davis Centre 1302
11:15-12:00	Industry Panel	Davis Centre 1302
12:00-13:30	MITACS Awards Luncheon	Festival Room, South Campus Hall
13:30-18:00	Excursion	St. Jacob's Village

Table of Contents

Letter from the Chair	I
Acknowledgements	II
Organizing Committee	IV
Reviewers	V
Paper Presentations	VII
Keynote: Information Distance From a Question to an Answer <i>Ming Li</i>	IX
Keynote: Role of Image Processing, Navigation and Robots in Image-guided Intervention <i>Nobuhiko Hata</i>	X
Keynote: Structural Variation Discovery in High Throughput Sequenced Genomes and Transcriptomes <i>S. Cenk Sahinalp</i>	XI
Optimising Locality-Sensitive Hashing on Sequences in the Context of Motif Finding. <i>Warren A. Cheung</i>	1
Vessel Tracking for Retina Images Based on Fuzzy Ant Colony Algorithm. <i>Sina Hooshyar, Rasoul Khayati and Reza Rezai</i>	6
Estimating heart movement and morphological changes during robot-assisted coronary artery bypass graft interventions. <i>Mathew A. Carias, Cristian A. Linte, Daniel S. Cho and Terry M. Peters</i>	11
SPiD: An SVM-Based Protein Discriminator for Outer Membrane Proteins. <i>Babak Alipanahi</i>	16
Application of a respiratory CT sequence's combined histogram to estimate intra-sequence lung's air volume variations. <i>Ali Sadeghi Naini, Rajni V. Patel and Abbas Samani</i>	21
A fast breast nonlinear elastography reconstruction technique using the Veronda-Westman model. <i>Mohammadhosein Amooshahi and Abbas Samani</i>	26
Evaluating a motor unit potential train using cluster validation methods. <i>Hossein Parsaei and Daniel W. Stashuk</i>	31

In silico study of the interaction of the Myelin Basic Protein C-terminal α -helical peptide with DMPC and mixed DMPC/DMPE lipid bilayers. <i>Kyrylo Bessonov</i>	36
The application of Quantum Tunneling Compound to sleep actigraphy. <i>Martin Bowyer, Ken Jones and Mila Kwaitkowska</i>	41
A Fast Technique of Tissue Biomechanical Analysis for Real-time Prostate Tissue Elasticity Reconstruction. <i>Seyed Reza Mousavi and Abbas Samani</i>	46
Finding Important Protein Motions in Solution Through Linear Dimensionality Reduction. <i>Krzysztof Borowski and Forbes Burkowski</i>	51
A Quick Introduction to Data Compression Through Learning. <i>Francisco Claude</i>	56
High gain lateral amorphous selenium (a-Se) detector for medical imaging. <i>Shiva Abbaszadeh, Kai Wang, Nicholas Allec, Feng Chen and Karim S. Karim</i>	61
An in silico mathematical model of the initiation of DNA replication. <i>Rohan D. Gidvani, Brendan J. McConkey, Bernard P. Duncker and Brian P. Ingalls</i>	65
On sample allocation in differential in-gel electrophoresis: twos a couple, threes a crowd?. <i>Owen Z. Woody, Lorna Deeth, Catherine Walden, Thomas D. Singer and Brendan J. McConkey</i>	70
Evaluation of New Parameters for Assessment of Stroke Impairment. <i>Kathrin Tyryshkin, Janice I. Glasgow and Stephen H. Scott</i>	71
Online pattern matching in sparse matrices and contact maps. <i>Robert Fraser</i>	76
Modulation transfer function of amorphous selenium digital x-ray detectors. <i>Yuan Fang, Nicholas Allec, Ling Guo and Karim S. Karim</i>	81
Using Dynamic Bayesian Networks to analyze genetic data. <i>Zhen Wang and Andrew Wong</i>	85
MeSHOP: MeSH Over-Representation Profiles for Summarising Biomedical Literature. <i>Warren A. Cheung, B.F. Francis Oullette, and Wyeth W. Wasserman</i>	90
Studying the Effect of Pulsatile Compulsory Blood Flow in Models of Stenotic Coronary Artery by Application of the Fluid-Structure Interaction. <i>Alireza Hashemifard and Nasser Fatouraee</i>	91
Aspects of Beta Amyloid Aggregation and Its Interaction with Acetylcholine Neurotransmitters and Alzheimers Disease. <i>Ibrahim Mustafa, Pu Chen, and Ali Elkamel</i>	92
Image Segmentation Using Varying Ellipses. <i>Farnoud Kazemzadeh, Thomas M. Haylock and Arsen R. Hajian</i>	93

Segmentation of Carotid Artery from Three-dimensional Ultrasound Images Using Active Contours. <i>Eranga Ukwatta, Joseph Awad, Aaron D. Ward, Adam Krasinski and Aaron Fenster</i>	94
Quantification of prostate deformation due to needle insertion during TRUS-guided biopsy. <i>Tharindu De Silva, Aaron Fenster, Jagath Samarabandu and Aaron D. Ward</i>	95
Cross-laboratory comparison of human post-mortem brain expression profiling data. <i>Meeta Mistry, Kelsey Hamer, and Paul Pavlidis</i>	96
Computer-Controlled Dynamic Human Gastrointestinal Model: In Vitro Exploration Of The Human Gut Microbiota. <i>Rodes L., Tomaro-Duchesneau C., Coussa-Charley M., Martoni C., Bhathena J., Prakash S.</i>	97
A real-time biomechanics analysis method for multifocal breast cancer assessment. <i>Shadi Shavakh, Aaron Fenster and Abbas Samani</i>	98
Fractal Time-Series Analysis of Reaching Motion in Stroke Affected Arms. <i>Tanny F. King, Kathrin Tyryshkin and Janice I. Glasgow</i>	99
Topology-aware closest point cortical thickness. <i>Eli Gibson and M. Faisal Beg</i>	100
Design and Implementation of a 3D Ultrasound System for Image Guided Liver Interventions. <i>Hamid R. Sadeghi-Neshat, Shi Sherebrin, Lori Gardi and Aaron Fenster</i>	101
Applying Normal Mode Analysis to the Conserved Patterns of Cytochrome C. <i>En-Shiun Annie Lee</i>	102
Monte Carlo simulation of amorphous Selenium digital x-ray detectors: spatial dependence of energy absorption. <i>Yuan Fang, Nicholas Allec and Karim S. Karim</i>	103
Reconstruction of Needle Tracts from Fluoroscopy in Prostate Brachytherapy. <i>Lauren E. Gordon, Ehsan Dehghan, Septimiu E. Salcudean and Gabor Fichtinger</i>	104

Letter from the Chair

Welcome to CSCBCE 2010, the 5th Canadian Student Conference on Biomedical Computing and Engineering. It has been a great pleasure to continue the tradition envisioned and initiated by my Master's degree supervisor Janice Glasgow. There have been wonderful groups of people associated with this conference each year, and I'm honoured to have been able to be a part of the group to organize the conference this year. In particular, I would like to express my gratitude to all of the organizing committee, the reviewers, the keynote speakers and tutorial instructors, the people who have helped promote the conference across Canada, and the sponsors who have supported us so generously.

The reception was tremendous this year, and it is great to see that there is a strong level of support for the conference amongst the students in these fields. CSCBCE has facilitated the establishment of a community of Canadian students who work in these areas, and it allows for the exchange of ideas between communities who may not traditionally share a venue.

Finally, I would like to extend my gratitude to you for attending our conference, and for contributing to our success. I hope that we may meet at a CSCBCE in the future!

Sincerely,

Bob Fraser
Chair, CSCBCE 2010

Acknowledgements

We would like to thank many individuals, groups and organizations for their support and involvement with CSCBCE 2010. A great number of people contributed their time, expertise, and resources:

Keynote Speakers

Nobuhiko Hata, Associate Professor of Radiology, Harvard Medical School

Ming Li, Canada Research Chair in Bioinformatics, University of Waterloo

S. Cenk Sahinalp, Canada Research Chair in Computational Genomics, Simon Fraser University

Tutorial Instructors

Shihyen Chen, University of Western Ontario

Xin Gao, Lane Fellow at School of Computer Science, Carnegie Mellon University

Mehdi Moradi, Department of Electrical and Computer Engineering, University of British Columbia

Special Thanks

Andrea Buddin, Events Coordinator, MITACS Inc.

Warren Cheung, Co-chair, CSCBC 2009

Vera Korody, Special Events/Operations Coordinator, Institute for Computer Research

Sherryl DiCiccio, Administrative Officer, David R. Cheriton School of CS, University of Waterloo

Peter Forsyth, Professor, David R. Cheriton School of Computer Science, University of Waterloo

Mark Giesbrecht, Associate Director, David R. Cheriton School of Computer Science, University of Waterloo

Lena Hussain, Coordinator, MITACS Step Program

Helen Jardine, Secretary, David R. Cheriton School of Computer Science, University of Waterloo

Mehdi Moradi, Chair, CSCBC 2006

Olena Morozova, Co-chair, CSCBC 2009

M. Tamer Özsu, Director, David R. Cheriton School of Computer Science, University of Waterloo

Anne Turnbull, Financial Officer, David R. Cheriton School of Computer Science, University of Waterloo

Platinum Sponsors

MITACS

University of Waterloo Graduate Student Endowment Fund (GSEF)

Silver Sponsors

David R. Cheriton School of Computer Science, University of Waterloo

Genome Canada

Bronze Sponsors

Mathematics Endowment Fund (MEF), University of Waterloo

Institute for Computer Research (ICR), University of Waterloo

Canadian Bioinformatics Workshops

Organizing Committee

Robert Fraser	Chair
Babak Alipanahi	Technical Coordinator
Francisco Claude	Webmaster
Salam Gabran	Organizing Committee
Natalie Fox	Delegate Coordinator
Chantelle Gomes	Undergraduate Publicity Coordinator
Richard Jang	Treasurer & Scientific Program Coordinator
Anna Merkoullovitch	Delegate Coordinator
Patrick Nicholson	Publications Coordinator
Alejandro Salinger	Logistics Coordinator
Somayyeh Zangoeei	Publicity Coordinator

Bin Ma	Faculty Advisor
--------	-----------------

Rosanne Borja	Volunteer
Jasmine Chan	Volunteer
Sarah Chan	Volunteer
James Gleeson	Volunteer
En-Shiun Annie Lee	Volunteer
Cristina Maiocco	Volunteer
Julie Moon	Volunteer
Plinio Morita	Volunteer

Reviewers

We graciously acknowledge the efforts of our reviewers to provide constructive feedback for their peers.

Mariam Afshin

*Department of Biomedical Engineering
University of Western Ontario, Canada*

Mehrdad Fahimnia

*Department of Electrical and Computer Engineering
University of Waterloo, Canada*

Babak Alipanahi

*School of Computer Science
University of Waterloo, Canada*

Yuan Fang

*Department of Electrical and Computer Engineering
University of Waterloo, Canada*

Ali Asadian

*CSTAR, London Health Sciences and ECE Department
University of Western Ontario, Canada*

Carol Fung

*David R. Cheriton School of Computer Science
University of Waterloo, Canada*

Ahmad Ashoori

*Department of Electrical and Computer Engineering
University of British Columbia, Canada*

Salam Gabran

*Department of Electrical and Computer Engineering
University of Waterloo, Canada*

Mahdi Azizian

*Department of Electrical and Computer Engineering
University of Western Ontario, Canada*

Sahar Ghanavati

*Department of Electrical and Computer Engineering
Queen's University, Canada*

Parisa Behnamfar

*Electrical and Computer Engineering Department
University of British Columbia, Canada*

Mireille Gomes

*Systems Biology Doctoral Training Centre
University of Oxford, Canada*

Malay Bhattacharyya

*Machine Intelligence Unit
Indian Statistical Institute, Kolkata, India*

Richard Jang

*David R. Cheriton School of Computer Science
University of Waterloo, Canada*

Rodrigo Cánovas

*Department of Computer Science
University of Chile, Chile*

Susana Ladra

*Department of Computer Science
University of A Coruña, Spain*

Warren A. Cheung

*Bioinformatics Program
University of British Columbia, Canada*

En-Shiun Annie Lee

*Department of System Design Engineering
University of Waterloo, Canada*

Francisco Claude

*David R. Cheriton School of Computer Science
University of Waterloo, Canada*

Cristian A. Linte

*Imaging Research Laboratories
Robarts Research Institute, Canada*

Meeta Mistry
Bioinformatics Program
University of British Columbia, Canada

Kathrin Tyryshkin
School of Computing
Queen's University, Canada

Patrick Nicholson
David R. Cheriton School of Computer Science
University of Waterloo, Canada

Ismael A. Vergara
Molecular Biology and Biochemistry Department
Simon Fraser University, Canada

Kyle Nishiyama
Biomedical Engineering
University of Calgary, Canada

Daniel Valenzuela
Department of Computer Science
University of Chile, Chile

Yves Pauchard
Department of Electrical and Computer Engineering
University of Calgary, Canada

Helen Xu
School of Computing
Queen's University, Canada

HamidReza SadeghiNeshat
Biomedical Engineering Graduate Program
University of Western Ontario, Canada

Somayyeh Zangooei
David R. Cheriton School of Computer Science
University of Waterloo, Canada

Alejandro Salinger
David R. Cheriton School of Computer Science
University of Waterloo, Canada

Paper Presentations

Friday Morning, Track A (Bioinformatics)

10:15-10:35	Optimising Locality-Sensitive Hashing on Sequences in the Context of Motif Finding <i>Warren Cheung</i>
10:35-10:55	SPiD: An SVM-Based Protein Discriminator for Outer Membrane Proteins <i>Babak Alipanahi Ramandi and Ming Li</i>
10:55-11:15	Finding Important Protein Motions in Solution Through Linear Dimensionality Reduction <i>Krzysztof Borowski and Forbes Burkowski</i>
11:15-11:35	Online Pattern Matching in Sparse Matrices and Contact Maps <i>Robert Fraser</i>
11:35-11:55	Using Dynamic Bayesian Networks to Analyze Genetic Data <i>Zhen Wang and Andrew Wong</i>

Friday Morning, Track B (Biomedical Engineering)

10:15-10:35	Estimating Heart Movement and Morphological Changes During Robot-Assisted Coronary Artery Bypass Graft Interventions <i>Mathew Anibal Fortes Carias, Cristian Linte, Daniel Cho and Terry Peters</i>
10:35-10:55	A Fast Breast Nonlinear Elastography Reconstruction Technique Using the Veronda-Westman Model <i>Mohammadhosein Amooshahi and Abbas Samani</i>
10:55-11:15	Evaluating a Motor Unit Potential Train Using Cluster Validation Methods <i>Hossein Parsaei and Daniel Stashuk</i>
11:15-11:35	The Application of Quantum Tunneling Compound to Sleep Actigraphy <i>Ken Jones, Martin Bowyer and Mila Kwaitkowska</i>
11:35-11:55	A Technique for Tissue Biomechanical Analysis for Real-time Prostate Tissue Elasticity Reconstruction <i>Seyed Reza Mousavi and Abbas Samani</i>

Friday Afternoon, Track A (Biomedical Science)

2:45-3:05	In Silico Study of the Myelin Basic Protein C-terminal alpha-Helical Peptide in DMPC and DMPE Lipid Bilayers to Further Multiple Sclerosis Research <i>Kyrylo Bessonov</i>
3:05-3:25	An in silico mathematical model of the initiation of DNA replication <i>Rohan D. Gidvani, Brendan J. McConkey, Bernard P. Duncker and Brian P. Ingalls</i>
3:25-3:45	On sample allocation in differential in-gel electrophoresis: two's a couple, three's a crowd? <i>Owen Woody, Lorna Deeth, Catherine Walden, Thomas Singer and Brendan McConkey</i>
3:45-4:05	Evaluation of New Parameters for Assessment of Stroke Impairment <i>Kathrin Tyryshkin, Janice Glasgow and Stephen Scott</i>
4:05-4:25	A Quick Introduction to Data Compression Through Learning <i>Francisco Claude</i>

Friday Afternoon, Track B (Medical Imaging)

2:45-3:05	Vessel Tracking for Retina Images Based on Fuzzy Ant Colony Algorithm <i>Sina Hooshyar and Rasoul Khayati</i>
3:05-3:25	Application of a respiratory CT sequence's combined histogram to estimate intra-sequence lung's air volume variations <i>Ali Sadeghi Naini, Rajni V. Patel and Abbas Samani</i>
3:25-3:45	High gain lateral amorphous selenium (a-Se) detector for medical imaging <i>Shiva Abbaszadeh, Kai Wang, Nicholas Allec, Feng Chen and Karim Karim</i>
3:45-4:05	Modulation transfer function of amorphous selenium digital x-ray detectors <i>Yuan Fang, Nicholas Allec, Ling Guo and Karim Karim</i>

Information Distance From a Question to an Answer

Ming Li^{a*}

^aCanada Research Chair in Bioinformatics
University Professor
David R. Cheriton School of Computer Science,
University of Waterloo,
Waterloo, ON, Canada, N2L 3G1

We know how to measure the distance from Toronto to Amsterdam. However, do you know how to measure the distance between two information carrying entities? For example: two genomes, two music scores, two programs, two articles, two emails, or from a question to an answer? Furthermore, such a distance measure must be application-independent, must be universal in the sense it is provably better than all other distances, and must be applicable.

From a simple and accepted assumption in thermodynamics, we have developed such a theory. I will present this theory and its applications. In particular, we will present a new application of the theory: a question answering system.

Biography

Ming Li is a Canada Research Chair in Bioinformatics and a University Professor at the University of Waterloo. He is a fellow of the Royal Society of Canada, ACM, and IEEE. He is a recipient of Canada's E.W.R. Steacie Fellowship Award in 1996, the 2001 Killam Fellowship, and the Killam Prize in 2010. Together with Paul Vitanyi they have pioneered the applications of Kolmogorov complexity and co-authored the book "An Introduction to Kolmogorov Complexity and Its Applications". In particular, his work on information distance and normalized information distance has found many applications in document comparison, genome evolution, time series analysis, as well as Question and Answer search engine on the internet. He is a co-managing editor of Journal of Bioinformatics and Computational Biology.

*E-mail: mli@uwaterloo.ca

Role of Image Processing, Navigation and Robots in Image-guided Intervention

Nobuhiko Hata^{a*}

^aAssociate Professor of Radiology, Harvard Medical School
Technical Director, Image Guided Therapy Program, Brigham and Women's Hospital
Director, Surgical Navigation and Robotics Laboratory
L1-050, Department of Radiology, Brigham and Women's Hospital
75 Francis St., Boston, MA 02115

Image-guided surgery is a promising method of cure to achieve minimal trauma, fast patient recovery, and reduction of clinical cost. Images used for navigation can be either pre-operative diagnostic images and/or intra-operative images. We have also observed the prevalence of surgical robots in clinics in the past decade giving an impact on medical care. The aim of this presentation is to discuss the new dimension of minimally invasive surgery we can explore by integrating intra- and pre-operative imaging and surgical robots. The robot presented in my talk can compensate for the motion of the organs and guide the precision surgery by using intra-operative images as a digital map for robot control, without which we cannot perform image-guided intervention of dynamically moving organs. The need for image-guided robotics is further highlighted in our long-term clinical goal in the Brigham and Women's Hospital, to perform therapies in contemporary high-field, closed-bore MRI scanners, 3D Ultrasound, CT, and/or PET/CT that provide better delineation of disease lesion, and that becoming to be prevalent in hospitals and clinics worldwide. I will present our pilot studies from MR-guided ablation and needle therapies where we studied tissue-needle interaction, biomechanical modeling of tissue deformation, and merit of needle guidance by a MR-compatible robot. Those studies will eventually migrate into our motion control method of a close-bore MRI-compatible robot that drives therapy needles toward the targets under intra-operative image guidance and control. For more information about my talk and exciting research opportunities at the Image Guided Therapy Program at Brigham and Women's Hospital and Harvard Medical School, visit www.snrlab.org , www.ncigt.org .

Biography

Nobuhiko Hata was born in Kobe, Japan. He received the B.E. degree in precision machinery engineering in 1993 from School of Engineering, The University of Tokyo, Tokyo, Japan, and the M.E. and the Doctor of Engineering degrees in precision machinery engineering in 1995 and 1998 respectively, both from Graduate School of Engineering, The University of Tokyo, Tokyo, Japan. He is currently an Associate Professor of Radiology, Harvard Medical School and Technical Director of Image Guided Therapy Program, Brigham and Women's Hospital. He started his career at Brigham and Women's Hospital initially as a research fellow in 1995, then became Instructor of Radiology in 2000 and Assistant Professor of Radiology in 2005, all at Department of Radiology. In 2008, He founded a research group called Surgical Navigation and Robotics Laboratory, under Image Guided Therapy Program. In the Image Guided Therapy Program and Surgical Navigation and Robotics Laboratory, he continues to work on medical image processing and robotics in image-guided surgery. His major achievements include neurosurgical navigation combined with ultrasound imaging, surgical robotics for magnetic resonance images, and motion-adaptable surgical robotics for image-guided therapy. More importantly, he developed key technology in many "the first" therapy in MRI-guided therapy; MR-guided prostate biopsy, MR-guided laser ablation therapy of brain tumor, and MR-guided microwave ablation therapy of liver tumor. In total, he has co-authored 49 original articles and has been involved in 11 federal and non-federal grants during the course of his research career.

*E-mail: hata@bwh.harvard.edu

Structural Variation Discovery in High Throughput Sequenced Genomes and Transcriptomes

S. Cenk Sahinalp^{a*}

^aDirector, SFU Lab For Computational Biology

Professor of Computing Science

Associate Faculty, Department of Molecular Biology and Biochemistry

Canada Research Chair in Computational Genomics

Michael Smith Foundation for Health Research Scholar

School of Computing Science, Simon Fraser University

8888 University Drive, Burnaby BC, V5A 1S6

Recent studies show that along with single nucleotide polymorphisms and small indels, larger structural variants contribute significantly to human genetic diversity. The realization of new ultra-high-throughput sequencing platforms now makes it feasible to detect the full spectrum of genomic variation among many individual genomes, including cancer patients and others suffering from diseases of genomic origin. Conventional algorithms for identifying structural variation (SV) have not been designed to handle the short read lengths and the errors implied by the “next-gen” sequencing (NGS) technologies. In this talk we will describe combinatorial formulations for the SV detection between a reference genome and a high throughput paired-end sequenced individual genome. We will provide efficient algorithms for each of the formulations we give, which all turn out to be fast and quite reliable; they are also applicable to all next-gen sequencing methods and traditional capillary sequencing technology.

Biography

Cenk Sahinalp is a Professor of Computing Science at Simon Fraser University, Burnaby BC, an associate faculty at the Department of Molecular Biology and Biochemistry and a visiting scientist at the Department of Genome Sciences, University of Washington. His research focuses on problems in sequence alignment, search and comparison, biomolecular sequence analysis with emphasis on structural variation detection through the use of high throughput sequencing, RNA structure and interaction prediction, biomolecular network analysis and small molecule bioinformatics. He is a Canada Research Chair, a Michael Smith Foundation Scholar and has been a recipient of an NSF Career Award in theoretical computer science. He has served /will serve as the PC chair of the Combinatorial Pattern Matching (CPM) Conference in 2004, the general chair of the RECOMB Conference in 2011, the area chair on sequence analysis, next-gen sequencing, RNA structure prediction, etc. for a number of conferences including ISMB and PSB.

*E-mail: cenk@cs.sfu.ca

Optimising Locality-Sensitive Hashing on Sequences in the Context of Motif Finding

Warren A. Cheung^{a*}

^aBioinformatics Program, Centre for Molecular Medicine and Therapeutics
University of British Columbia, Vancouver, BC

ABSTRACT

Locality-sensitive hashing has been applied in several problems in bioinformatics to quickly search large sequences by examining the n -grams of the sequences. We observe in these application a bias in the random generation of hashes and define an effective equivalence for hashes that generate nearly identical results. Methods that address these two points demonstrate an improvement to the rate at which unique hashes are generated, especially when many hashes are generated. These methods can be easily integrated with existing applications of locality-sensitive hashing, resulting in improved performance by reducing the time algorithms spend revisiting duplicate hashes and eliminating search biases.

Keywords: locality-sensitive hashing, projection, motif-finding, n -grams, pattern discovery, similarity search

1 Introduction

The random projection method of locality sensitive hashing is a stochastic method of dimensionality reduction. This technique has been successfully applied to many problems involving identification of patterns, such as motif-finding and local sequence alignment. This paper discusses optimisations in the context of the motif finding problem originally proposed by Pevzner and Sze[1], with the application of locality sensitive hashing method PROJECTION described by Buhler and Tompa[2], which uses random projection as a efficient method for finding common semi-conserved motifs of a set length without gaps from a set of sequences. We demonstrate that careful generation around a set of constraints can result in a significantly reduced search space with effectively equivalent results.

Previously, Raphael et al. developed UNIFORM PROJECTION[3], another improvement of PROJECTION that samples the projection space in a more uniform manner. We note that the optimizations discussed here are orthogonal from those, and further improvement to performance may be possible by combining with their work, such as previously achieved by Wang and Yang in UPNT[4] combining uniform projection with unpublished work on neighbourhood thresholding[5]. As well, the same analysis generalises to other locality-sensitive hashing-based methods, especially those involved in n -gram analysis, such as large-scale sequence comparison[6].

1.1 Classic Random Projection Algorithm

In the motif finding problem originally described by Pevzner and Sze[1], let S be a set of t input sequences with length n , and a motif of known length l with at most d mutations be planted in each of the t sequences. The purpose of the projection algorithm is to extract the t planted instances of the motif from S . For each sequence S_i in S , a sliding window of length l is run over S_i to generate a set of $n - l + 1$ “ l -mers” for S_i . Characters in a string shall be referred with 0 as the first character of the string. Let $s_{i,j}$ refer to the j^{th} symbol in S_i , where $0 \leq j < n$. Likewise, let $L_{i,j}$ refer the l -mer from S_i starting with symbol $s_{i,j}$, such that $L_{i,j} = s_{i,j}s_{i,j+1}\dots s_{i,j+l-1}$. Let L be the set of all l -mers.

A projection P is a set of k positions, each position representing one of the l positions $\{0, 1, \dots, l-1\}$. In random projection, each position is chosen by selecting one of the l positions uniformly at random with replacement and adding the position selected to P . This process is repeated k times. As the positions are chosen with replacement, a projection can specify from 1 to k distinct positions.

Given a projection P , each of the l -mers from S are hashed into a “bucket” by the bases of the l -mer at the positions specified by P . For a projection P of positions $\{p_1, p_2, \dots, p_k\}$ and l -mer $L_{i,j}$, let $P(L_{i,j}) = s_{i,j+p_1}s_{i,j+p_2}\dots s_{i,j+p_k}$. A bucket $B_{P,x}$ is the set of all l -mers $L_{i,j}$ where $P(L_{i,j}) = x$. A bucket containing an unusually large number of l -mers has an elevated chance of being enriched for a motif and is used as the basis for a “refinement” step to extract a motif

*E-mail: wcheung@cmmmt.ubc.ca, Telephone: +1(604)875-2345 ext. 7947.

via an EM-based approach (see Buhler and Tompa[2] for more details). We shall demonstrate that the method of choosing positions uniformly at random with replacement results in biased selection of projections.

2 Selection Bias

The current selection method introduces a bias in the selection of potential projections towards those that can be generated in multiple ways if the positions are selected uniformly at random with replacement. Naively, selecting from n possible positions k times results in n^k possible projections. However, it is obvious that the order of positions chosen for projection does not matter. Take two projections P and Q which are identical other than the order of the projected positions. Let f be a function that reorders the positions in P to the positions of Q . We can then note that the number of l -mers in bucket $B_{P,x}$ is the same as bucket $B_{Q,f(x)}$.

Definition Projection P is *equivalent* to Q if there exists a correspondence f such that for any set of sequences, for every l -mer $L_{i,j}$, $f(P(L_{i,j})) = Q(L_{i,j})$. P and Q are *exactly equivalent* if P is equivalent to Q and vice versa.

Definition Projection Q is *non-redundant* if for $Q = \{q_1, q_2, \dots, q_k\}$, for any q_i, q_j where $i \neq j$, $q_i \neq q_j$.

Under this definition of equivalence, it can be shown that any projection P is exactly equivalent to a non-redundant projection $Q - f$ in this case reorders the positions as needed with duplicate positions removed. Therefore, we shall use the following *standard form* for listing positions in projections: (i) the projected positions are in sorted order and (ii) no positions are duplicated. Note that all projections are equivalent to a projection in standard form, and no two different projections in standard form are equivalent. If only projections in standard form are considered, the search space of possible projections is reduced to $\sum_{i=1}^k \binom{n}{i}$.

When selecting positions uniformly at random with replacement, selection bias occurs because some projections can be generated by more combinations of randomly selected positions than others. For example, for $k = 3$, the only way to randomly generate the projection $\{5\}$ is by selecting position 5 for all three random draws $(5, 5, 5)$. This yields the non-redundant projection $\{5\}$, and no other projection can be generated that is equivalent to this projection. However, the draws $(3, 1, 3)$, $(1, 3, 3)$, $(3, 3, 1)$, $(3, 1, 1)$, $(1, 3, 1)$ and $(1, 1, 3)$ all generate a non-redundant projection of $\{1, 3\}$, making it three times as likely to be generated as $\{5\}$.

The selection bias increases duplicate projections. As the projection algorithm is deterministic, duplicate projections cause redundant computation of already-evaluated results. Even when duplicate projections are not chosen, the projection choices are biased towards projections that have more ways to be generated. Such a bias makes the method less effective when the solution involves a less likely to be generated projection.

2.1 Avoiding Selection Bias

An efficient method to generate projections with equal frequency is to modify the random selection process. The new process selects k elements at random without replacement from the set $\{0, 1, \dots, l-1\} \cup \{\emptyset_0, \emptyset_1, \dots, \emptyset_{k-2}\}$. The $k-1$ elements \emptyset_i are *pseudo-positions* not in the set of original positions $\{0, 1, \dots, l-1\}$. Let S be the set of k selected elements. If S contains no pseudo-positions or $S \cap \{\emptyset_0, \emptyset_1, \dots, \emptyset_{k-2}\} = \{\emptyset_0, \emptyset_1, \dots, \emptyset_i\}$ for some i , S defines the projection $S \cap \{0, 1, \dots, l-1\}$. Otherwise, S is discarded and the selection process is restarted from the beginning.

This method is slightly more costly than sampling at random, as several restarts may be necessary before a valid projection is generated. However, it guarantees every potential projection is generated with equal frequency, as there a set of equivalent projections can only be generated in one manner.

3 Effective Equivalence of Projections

As the l -mers in projection are generated by sliding a window across the input sequences, sequence positions are always shared between neighbouring l -mers. A position $s_{i,j}$ appears in up to l of the l -mers: $L_{i,j-l+1}, L_{i,j-l+2}, \dots, L_{i,j}$. One side effect of this effect is that many combinations of projections P and l -mer $L_{i,j}$ will result in the same sequence $P(L_{i,j})$. For example, let projections P and Q be $P = \{2, 4\}$ and $Q = \{1, 3\}$. For $1 \leq j \leq n-1$, $P(L_{i,j}) = Q(L_{i+1,j+1})$. This motivates a more relaxed definition of equivalence between two projections.

Definition A mapping Δ from $L' \rightarrow L$ is a *shift* of L if $\Delta(L_{i,j}) = L_{i,j-\delta}$ and $\delta \geq 0$, where L' is the set of all l -mers $L_{i,j} \in L, j \geq \delta$.

The shift Δ maps nearly all the l -mers by a constant shift along the sequence to a nearby l -mer. However, a constant number of l -mers at the start of each sequence are unable to be mapped.

Definition Projection P is *effectively equivalent* to Q if there exists a shift Δ such that for any set of sequences, for every l -mer mapped by Δ , $P(\Delta(L_{i,j})) = Q(L_{i,j})$.

This relaxes our definition of equivalence, allowing P and Q to be equivalent as long as for the l -mers mapped by Δ (all l -mers except a small constant number δ), a projection $Q(L_{i,j})$ can be mapped to the projection $P(\Delta(L_{i,j}))$. Note when we have a shift $\delta = 0$, we have exact equivalence, therefore effective equivalence includes exact equivalence.

One specific case of such an effective equivalence is that when $0 \notin P$, P is effectively equivalent to the projection Q where $Q = \{x | (x + \min(P)) \in P\}$. This case corresponds to a shift Δ where $\delta = \min(P)$.

Proof. By contradiction, assume there exists a projection P , $0 \notin P$, where P is not effectively equivalent to Q , where $q_i = p_i - \min(P)$ and $\Delta(L_{i,j}) = L_{i,j-\min(P)}$. Then there must exist an l -mer $L_{i,j}$ such that $P(\Delta(L_{i,j})) \neq Q(L_{i,j})$. However,

$$\begin{aligned} P(\Delta(L_{i,j})) &= P(L_{i,j-\min(P)}) \\ &= s_{i,j+p_1-\min(P)} s_{i,j+p_2-\min(P)} \cdots s_{i,j+p_k-1-\min(P)} \\ &= s_{i,j+q_1} s_{i,j+q_2} \cdots s_{i,j+q_k-1} = Q(L_{i,j}). \end{aligned}$$

Therefore, no such l -mer $L_{i,j}$ exists and therefore P is effectively equivalent to Q . \square

Therefore, any projection P is effectively equivalent to a projection Q in standard form, where $\min(Q) = 0$. This reduces the space of possible projections where no two projections are effectively equivalent to at most $\sum_{i=1}^{k-1} \binom{n}{i}$.

Note that the generation of effectively equivalent projections is also a source of bias in the generation of projections, as some projections are effectively equivalent to more projections than others. For example, when $l = 3$, the projection $\{3, 4, 6\}$ is also effectively equivalent to the projections $\{2, 3, 5\}$, $\{1, 2, 4\}$ and $\{0, 1, 3\}$ whereas the projection $\{0, 2, 6\}$ has no other effectively equivalent projection and so is four times less likely to be generated (assuming the unbiased generation of projections described in Section 2.1).

3.1 Avoiding Effectively Equivalent Projections

To avoid generating effectively equivalent projections, it is simply necessary to ensure that the position 0 is part of the projection generated. The simplest solution is what we shall refer to as 0-base Projections, which is to ensure that the first position selected for any projection is position 0. The remaining positions can be selected as before, with the only other change being that position 0 be excluded from the future draws (as it has already been generated).

4 Results

To demonstrate the effect of these optimisations, the number of unique projections, or hashes, generated using the original basic hashing method (as seen in UPNT[4]) was compared to the number of unique hashes generated by the unbiased hashing method (See Section 2.1). Experiments were performed 100 times for each condition. The one-tailed unpaired t -test assuming non-equal variance was used to compute p -values. The test condition was projecting up to $k = 7$ positions of $l = 15$ l -mers. This test condition replicates the parameters for PROJECTION when solving the classic $(15 - 4)$ ($l - d$) motif-finding problem[2]. Timing results used a Python (<http://www.python.org>) implementation¹ and the internal `timeit.Timer` class, running the hashing function 1000 times.

4.1 Uniqueness of Hash Generation Methods

Figure 1 compares the number of unique hashes – hashes that are not equivalent. The unbiased hash function generates more unique hashes than the basic hash function, markedly so as the number of iterations increases. However, the difference is significant even at the smallest number of iterations tested ($p < 1.24 \times 10^{-7}$ at 500 iterations). A five percent improvement in the number of unique hashes generated is seen by 7500 iterations.

Even more striking is the number of hashes which are effectively unique – hashes which are not effectively equivalent. As the number of iterations increases, many of the hashes generated are effectively equivalent to a previously generated hash. At the extreme, after 25000 iterations, an average of 12819 unique hashes were generated by the unbiased hash generation method, but only 5966 of these were effectively unique – not effectively equivalent to a previously generated hash. To compare, 11659 unique hashes were generated on average by the original hash generation method, of which 5428 were effectively unique. The difference in effectively unique hashes generated is significant even at the smallest number of iterations tested ($p < 3.4 \times 10^{-23}$ at 500 iterations). Five percent improvement was seen by 2000 iterations (See Figure 3).

Figure 2 shows the effect of using the 0-base method on generating effectively unique hashes. The number of effectively equivalent hashes is increased substantially as effectively equivalent hashes are no longer generated. 0-base hash generation dominates the original hashing methods. As well, using 0-base, unbiased hash generation generates more effectively unique hashes than basic hash generation, even after the smallest number of iterations tested ($p < 7.06 \times 10^{-11}$ at 500 iterations). Using 0-base unbiased hashing has over five percent improvement over basic hashing at 500 iterations, rising to 26% by 10000 iterations (See Figure 3).

¹Source and raw data at <http://dnahelix.wikidot.com/locality-sensitive-hashing-for-motif-finding>

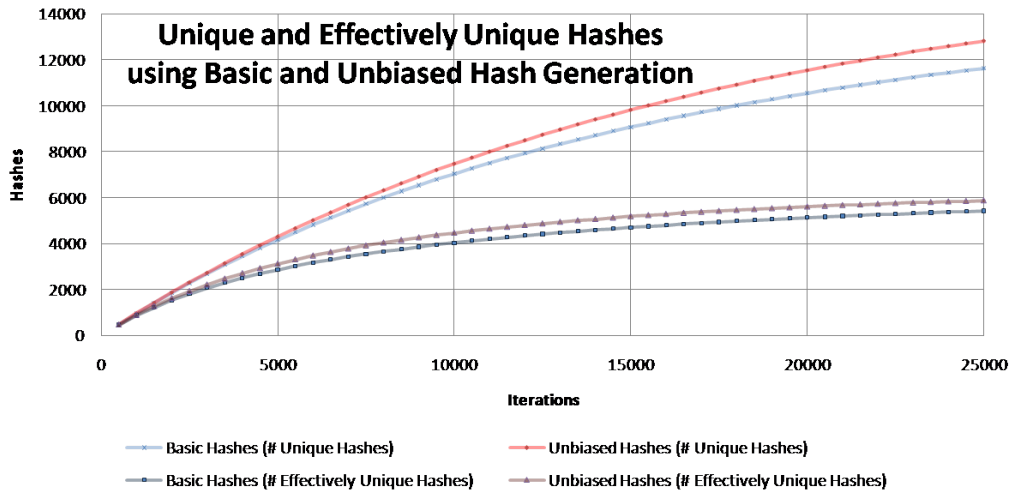


Figure 1. Average number of hashes that are unique and effectively unique over 100 trials. Unbiased hashing produces more hashes than basic hashing, however, the number of unique hashes is greatly reduced if we consider effective equivalence.

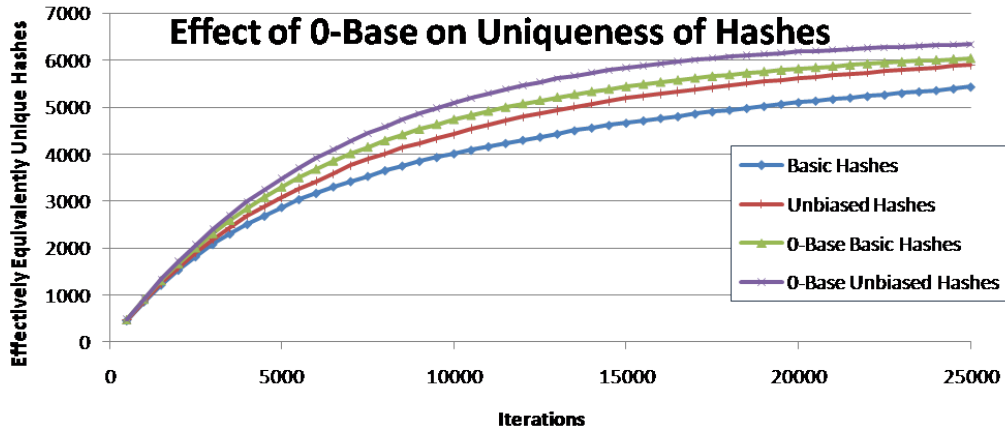


Figure 2. Average number of effectively unique hashes over 100 trials using unbiased and/or 0-base hashing. The unbiased hash generates more effectively unique hashes than the basic hash, and using 0-base increases this further.

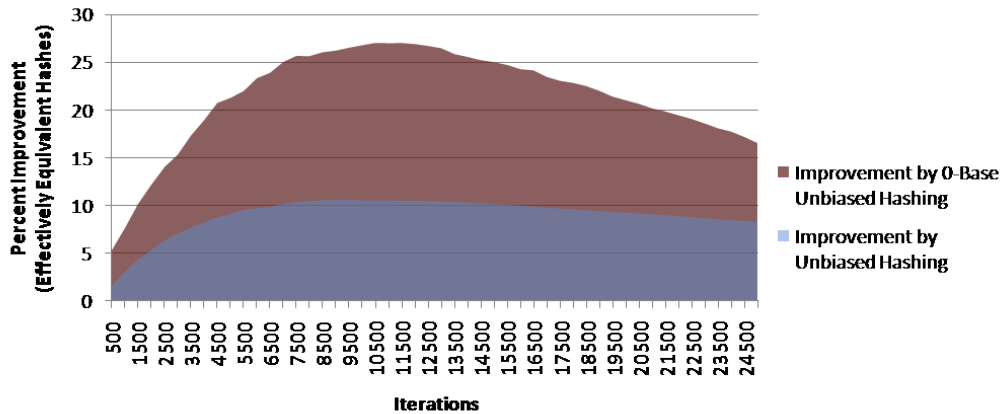


Figure 3. Percentage improvement on the number of effectively unique hashes, comparing unbiased hashing and unbiased with 0-base hashing over basic hashing. Using both types improvements, over 27% improvement was seen after 10500 iterations.

4.2 Runtime

The main caveat to using unbiased hash generation is the performance penalty incurred when rejecting hashes that are not in the desired format, which can happen with significant frequency. For our test case, the unbiased hash generation took on average 7.00×10^{-5} s to execute, whereas the basic hash generation took only 5.45×10^{-6} s. However, this is mitigated by the fact that selecting a hash function is a relatively short step in the motif finding algorithm, and would result in a net gain in cases where significantly more computation time is spent applying the hash function and analysing the result, as is generally the case. One refinement would be the pseudo-random generation of the hashes, which could guarantee no duplicates are generated. The application of uniform projection[3] or storing and rejecting all previously generated hashes could also mitigate this effect.

On the other hand, avoiding effectively equivalent projections actually simplifies generating the hash, as there is one fewer random trial to perform. The unbiased hash generation took slightly less time than before, on average 6.12×10^{-5} s. The basic hash generation was also only negligibly improved to 5.39×10^{-6} s.

5 Conclusion

We have demonstrated here two methods to reduce the search space of possible hashes for applications of locality-sensitive hashing in motif discovery. We note a bias in the random generation of hashes and define an effective equivalence for hashes that generate nearly identical results. Methods to address both these points are demonstrated to improve the uniqueness of hashes generated, especially when many hashes are generated.

Acknowledgements

This work was made possible due to funding by the CIHR/MSFHR Strategic Training Program in Bioinformatics, Natural Sciences and Engineering Research Council of Canada, the Michael Smith Foundation for Health Research, the Ontario Institute for Cancer Research and the University of British Columbia. Thanks to William S. Evans and Leon French for their input.

References

- [1] P. A. Pevzner and S.-H. Sze, "Combinatorial approaches to finding subtle signals in DNA sequences," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-00)*, pp. 269–278, AAAI Press, (Menlo Park, CA), Aug. 16–23 2000.
- [2] J. Buhler and M. Tompa, "Finding motifs using random projections," *Journal of Computational Biology* **9**, pp. 225–242, Apr. 2002.
- [3] B. Raphael, L.-T. Liu, and G. Varghese, "A uniform projection method for motif discovery in DNA sequences," *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **1**(2), pp. 91–94, 2004.
- [4] J. Wang and D. Yang, "UPNT: Uniform projection and neighbourhood thresholding method for motif discovery," *International Journal of Bioinformatics Research and Applications* **4**(1), pp. 96–106, 2008.
- [5] J. King, W. Cheung, and H. Hoos, "Neighbourhood thresholding for projection-based motif finding." Unpublished Manuscript.
- [6] J. Buhler, "Efficient large-scale sequence comparison by locality-sensitive hashing," *Bioinformatics* **17**(5), pp. 419–428, 2001.

Vessel Tracking for Retina Images Based on Fuzzy Ant Colony Algorithm

Sina Hooshyar^{a,1}, Rasoul Khayati^b and Reza Rezai^c

^{a,b,c}Department of Biomedical Engineering, Engineering Faculty, Shahed University, Khalije Fars Highway, 3319118651, Tehran, Iran

ABSTRACT

In this paper we present a novel fuzzy algorithm for vessel tracking in retina images. The main tools of this system are Ant Colony Optimization algorithm (ACO) and eigenvector analysis of Hessian matrix. ACO, inspired by food-searching behaviors of ants and performs well in discrete optimization, has been used for optimizing objective function of fuzzy C-means(FCM) model and clustering pixels into vessel and background clusters and Hessian matrix has been used for determining vessel direction in tracking process. Estimating full vessel parameters, overcoming initialization and profile modeling in related works and handling junction of vessels in retina image are the most important advantages of this method. Experiments and results of proposed algorithm in ocular fundus image show its good performance in vessel tracking and parameters estimating.

Keywords: Ant colony algorithm, Fuzzy clustering, Hessian matrix, Vessel tracking, Retina images

1 INTRODUCTION

The detection and measurement of blood vessels can be used to quantify the severity of disease, as part of the process of automated diagnosis of disease or in the assessment of the effect of therapy. Retinal blood vessels have been shown to change in diameter, branching angles or tortuosity as a result of a disease. Thus a reliable method of vessel segmentation would be valuable for the early detection and characterisation of morphological changes [1].

Vessel tracking is one of the common methods that are used in vessel segmentation. Most of vessel tracking methods begin from given initial points on the vessel and estimate the vessel width and orientation within a local region about the current point. Then a small step is taken along vessel direction and the procedure is repeated until stop conditions are satisfied. This method can calculate vessel centerline and diameter efficiently and provide a meaningful description of the vessel network. In the case of retina images, some works have been done according to this method that we can mention to [2-5]. In these papers, tracking was performed with respect to local information and tried to find maximum coincidence of vessels profile model.

In our retina processing system, we automatically initialize starting points from optic nerve in retina image and the profile that is centered in starting point is acquired at the direction normal to the direction of proper eigenvector of Hessian matrix in center point. Ant colony algorithm, which is a novel method to minimize objective function of FCM model and known as ACO-FCM, is used for detection of vessel and background regions along vessel profile. The position of real center is calculated by finding weighted mean of profile pixel positions with membership function of vessel cluster as their weights. The next center point is approximated by using of vessel direction and look-ahead distance which is proportional to vessel diameter.

The remaining of this paper is organized as follows: Section 2 provides a brief description of the Ant colony algorithm and its fuzzy application. Section 3 gives an overview of eigenvector analysis of Hessian matrix. Section 4 represents the proposed vessel tracking algorithm. In section 5, the experimental results are described and finally we express conclusions.

2 ACO-FCM CLUSTERING

Ant colony optimization was originally introduced by Dorigo and Maniezzo in 1996 [6]. In their original work, an optimization algorithm called *ant system* was introduced and applied to discrete optimization problems such as the traveling salesman problem (TPS). After this original article many other applications of ACO were reported [7].

¹ Corresponding author. E-mail: Sina.Hooshyar@gmail.com

The main idea in ACO is to mimic the behavior of real biological ants in search of food. The ants are able to efficiently find the shortest path from the nest to the food source and back. Ants deposit *pheromone* trails along their paths depending on the length of the trail (the shorter the trail, the more pheromones are deposited) and ant moves more or less randomly, but prefers locations with higher pheromone concentrations. The pheromones evaporate over time; hence the paths can be abandoned if they were not preferred during time. The ACO algorithm imitates these mechanisms by choosing solutions based on pheromones and updating pheromones according to the solution quality (and evaporation) [7]. More information about ant systems can be found in [8].

In traditional FCM model for clustering, a dataset $X = \{x_1, \dots, x_n\}$ is classified into $c \in \{2, \dots, n-1\}$ clusters that the following objective function must be minimized:

$$J(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m d_{ik}^2 \quad (1)$$

where U is the fuzzy partition matrix, $V = \{v_1, \dots, v_c\}$ is cluster centers vector, $c \geq 2$ is the number of clusters of final partition, n is the number of available data, the elements $u_{ik} \in [0, 1]$ of U represent the membership of data object x_k in cluster i , $m > 1$ is the *fuzzifier* that controls the fuzziness of the final partition, $d_{ik} = \|v_i - x_k\|$ is a distance metric between the data vector x_k and cluster center v_i .

The FCM clustering algorithm calculates partition matrix $U \in M_{fcm}$, where:

$$M_{fcm} = \left\{ U \in [0, 1]^{c \times n} \left| \sum_{i=1}^c u_{ik} = 1, k = 1, \dots, n, \sum_{k=1}^n u_{ik} > 0, i = 1, \dots, c \right. \right\}. \quad (2)$$

It can be shown that the necessary conditions for local minimum of objective function $J(U, V)$ are

$$v_i = \frac{\sum_{k=1}^n u_{ik}^m x_k}{\sum_{k=1}^n u_{ik}^m} \quad u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)}} \quad (3,4)$$

In 2005 Thomas A. Runkler [7] showed, by applying ACO-FCM algorithm to two types of databases, that FCM model optimized by ACO has better outcome than traditional optimization algorithm. In this algorithm each ant represents one of the data points $x_k \in X$ and assigns it to one of the C clusters based on a pheromone matrix $P \in R^{c \times n}$. Each entry in the pheromone matrix P represents one entry in the partition matrix U . The basic idea is to randomly produce fuzzy partition $U \in M_{fcm}$ whose expected value approximately corresponds to the normalized pheromone matrix P . This is done by adding Gaussian noise with variance σ to the normalized matrix P . After calculating objective function, pheromones matrix is updated according to evaporation rate and pheromone update function as following:

$$P_{ik} = P_{ik} \times (1 - \rho) + u_{ik} / (J(U, V) - J(U, V)_{min} + \varepsilon)^\alpha \quad (5)$$

where ρ is evaporation rate, $\varepsilon > 0$ and $\alpha > 1$ are user-specified parameters and $J(U, V)_{min}$ is minimum objective function which has ever been achieved. In any iteration that $J(U, V)$ is more than $J(U, V)_{min}$, P_{ik} is updated based on difference between $J(U, V)$ and $J(U, V)_{min}$ otherwise we have $J(U, V)_{min} = J(U, V)$ if $J(U, V)$ is less than $J(U, V)_{min}$, therefore, P_{ik} does not have very changes. The resulting ACO-FCM algorithm is summarized in Fig. 1.

3 EIGENANALYSIS OF THE HESSIAN MATRIX

Application of Hessian matrix to detect and analyze line-like structures has been investigated in many literatures and it is also used for segmentation and visualization of curvilinear structures in medical images [9,10]. The Hessian matrix is defined as:

$$H = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{bmatrix}. \quad (6)$$

Here, the second-order spatial derivative I_{ab} is calculated by convolution between the input image and scaled second-order derivative of Gaussian filter:

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \quad (7)$$

The eigenvalues and eigenvectors of Hessian matrix denote vessel's intensity and direction properties. Let λ_1 and λ_2 be the eigenvalues of Hessian matrix in given point as $|\lambda_1| \leq |\lambda_2|$ and v_1 and v_2 be the corresponding eigenvectors. It can be shown that vector v_1 is parallel to vessel axis-line while v_2 is perpendicular to that.

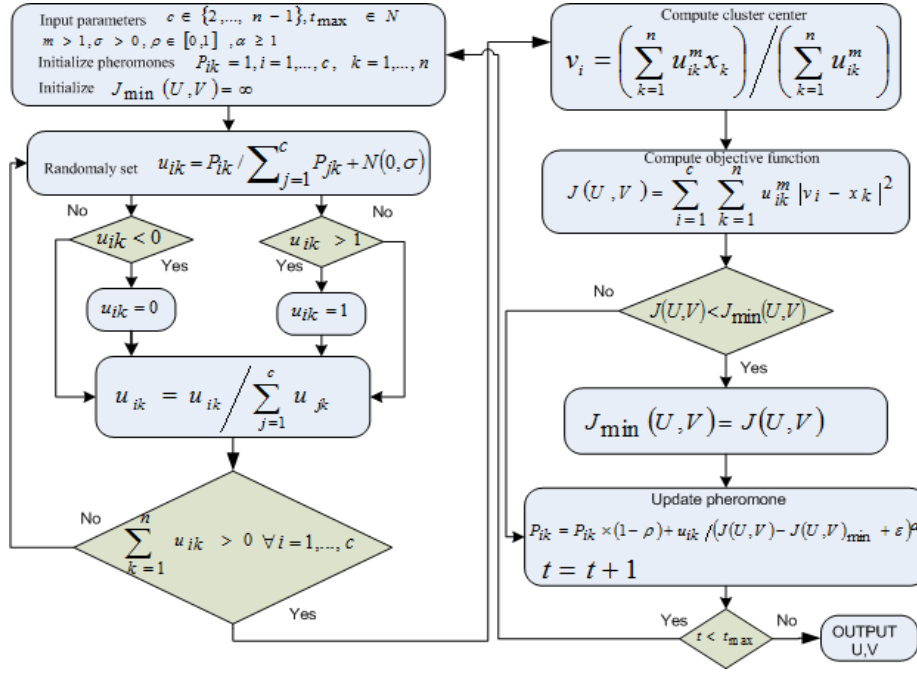


Figure 1. Flowchart of ACO-FCM algorithm

4 PROPOSED APPROACH

4.1 Initialization

There are some papers explaining how optic nerve can be found in retina images [11, 12]. These papers can be used to initialize algorithm. Having found the optic nerve, we form a sequence of points belonging to circle that bounds optic nerve. They are classified into vessel and background clusters by ACO-FCM algorithm. Each region in sequence that has more than three points with high membership degree in vessel cluster is considered as vessel candidate for starting the tracking process [5]. Center point of each region and its eigenvector of Hessian matrix in that point are defined as initial center pixel of vessel and its direction, respectively. False candidate points will be omitted within tracking algorithm.

4.2 Tracking process

The tracking will be started with initial points and its proper eigenvector of Hessian matrix. Let C_k be a pixel on vessel centerline in current iteration and v_k be eigenvector of Hessian matrix in center point that v_k is parallel to vessel direction. The location of center point in next iteration, C'_{k+1} , is calculated by:

$$C'_{k+1} = C_k + D_k v_k \quad (8)$$

where D_k is look-ahead distance parameter and it is proportional to vessel diameter in pervious iteration.

Centered at this position, a profile vector P is obtained by sampling of gray-scale values pixels along a scanline perpendicular to the direction of eigenvector in current center position v'_{k+1} . The length of profile is adapted to vessel diameter and is considered three times as much as vessel diameter in our work. The pixels gray-scale values of profile are classified into vessel and background clusters by ACO-FCM algorithm and center point of vessel is calculation by finding weighted mean of profile pixel position with membership function of vessel cluster as their weights:

$$\tilde{C}_{k+1} = \frac{\sum_{i=1}^n m_{vessel}(i) \times P(i)}{\sum_{i=1}^n m_{vessel}(i)} \quad (9)$$

where m_{vessel} denotes membership function of vessel cluster.

In order to compensate vessel changes in situations that it has high curvature and adjust center point between two edges, the profile P is obtained again with \tilde{C}_{k+1} and \tilde{v}_{k+1} as its center point and normal direction, respectively, that it is suggested by Sun [13] in 1989. The ACO-FCM classifies new profile and final center point is calculated by Equation (9). Right and left edge can be estimated as positions which have vessel and background

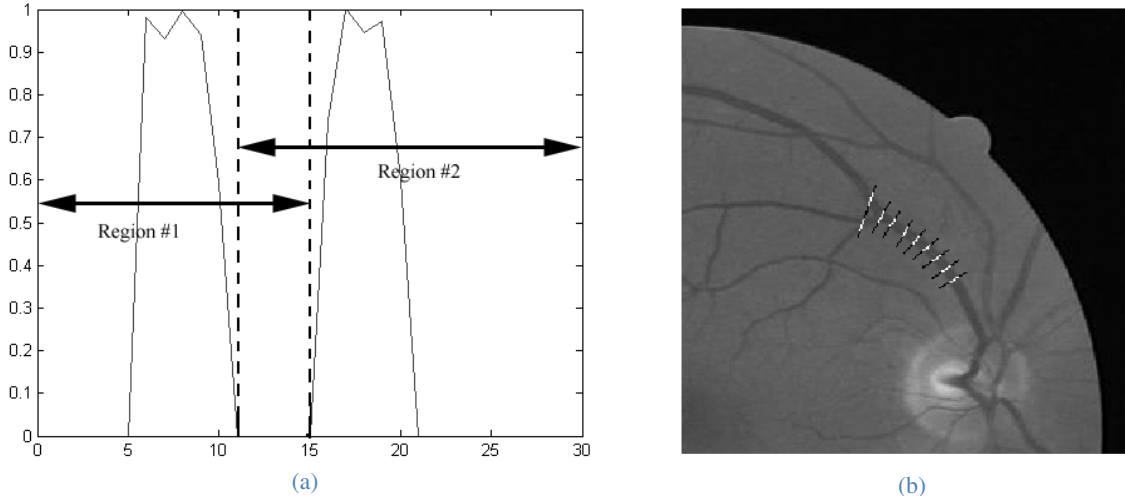


Figure 2. (a) The vessel membership function in bimodal state and two regions for calculating centers and (b) detection of junction corresponding to it.

membership functions almost equal in profile and vessel diameter is defined as distance between right and left edges.

4.3 Junctions

When algorithm confronts junctions in its tracking, the profile becomes bimodal. In these cases, profile is separated into two regions in order to choose better path as Fig. 2(a). After each part is classified by ACA-FCM, its center is calculated by Equation (9) and the center point which its eigenvector of Hessian matrix has less angle with vessel direction in last valid profile is selected as next center point to continue tracking and the other center point is stored to be processed later.

4.4 stopping criterion

When vessel diameter is less than a specified threshold or right and left edges become very close, the algorithm terminates and stores centers points, right and left edges and diameters as vessel attributes and processes another initial point.

5 RESULTS AND DISCUSSION

This algorithm has two groups of parameters. The first group is the ones belonging to ACO-FCM algorithm ($m, \sigma, \varepsilon, \alpha, \rho, t_{max}$) and the other is tracking parameters (look-ahead, profile size). The value of fuzzifier m for calculating the partitions is set to 2. This is the value that is usually used in the literature. σ is the variance of Gaussian noise added to the normalized pheromone matrix while $\varepsilon > 0$ and $\alpha > 1$ are user-specified parameters for updating pheromone matrix. An important parameter in ACO-FCM is evaporation rate (ρ). If ρ is chosen very high, pheromones matrix in Equation (5) will be affected by random numbers (u_{ik}) and worse results will be obtained. t_{max} is the iteration number of ACO-FCM and it must be high enough for decreasing objective function. The values of objective function versus the number of iteration have been shown in Fig. 3. In the second group, look-ahead distance influences computation time and it might miss several junction if it is selected large, hence, it is proportional to vessel diameter in pervious iteration. A fixed, large profile size would facilitate the detection of junction. However, in the case of vessels having small diameter the clustering algorithm would not provide valid cluster descriptions, therefore, the profile size is flexible and it is three times as much as last valid vessel diameter.

The vessel center and edge points of retina subimage extracted by algorithm are shown in Fig. 4. when $t_{max} = 1000, \rho = 0.25, \varepsilon = 0.01, \sigma = 0.001, m = 2$.

6 CONCLUSIONS

In this paper we have investigated the tracking of blood vessels in retina images. The proposed tracking scheme does not need any user interaction or any model for vessel profile. As well as it efficiently handles junctions of vessels in angiograms. The initial points are automatically obtained from optic nerve and ACO-FCM is used to classify pixel along vessel profile, which is normal to vessel orientation obtained by eigenanalysis of the Hessian matrix, into vessel and background cluster. In addition, vessel parameters such as centerlines, edge lines and

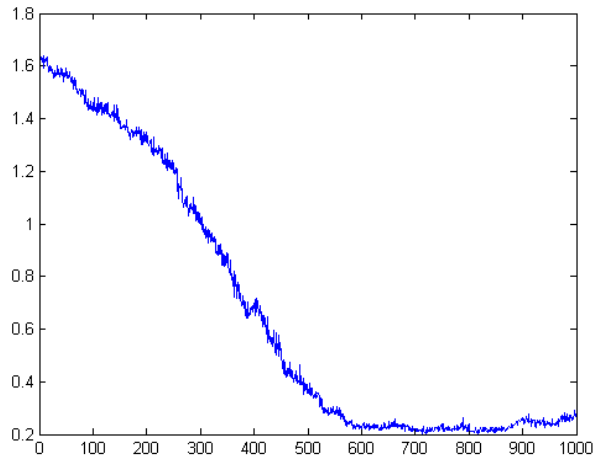


Figure 3. decreasing of objective function in one specific run of ACO-FCM

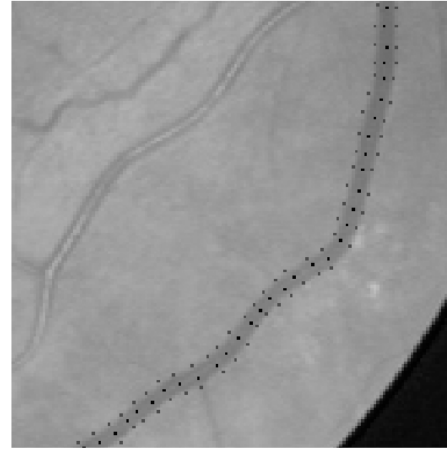


Figure 4. The main result of proposed algorithm

diameters are calculated by proposed algorithm. The results demonstrate the good performance of method in the whole tracking process and detecting more complete vessel network in the ocular fundus photographs.

References

- [1]. M. E. Martinez-Perez a, Alun D. Hughes, Simon A. Thom, Anil A. Bharath, Kim H. Parker, "Segmentation of blood vessels from red-free and fluorescein retinal images," In *Medical Image Analysis*, pages:47-61, 2007
- [2]. X. Gao, A. Bharath, A. Stanton, A. Hughes, N. Chapman, and S. Thom. "A method of vessel tracking for vessel diameter measurement on retinal images," In *ICIP01*, pages II: 881-884, 2001.
- [3]. M. Lalonde, L. Gagnon, and M.-C. Boucher. "Non-recursive paired tracking for vessel extraction from retinal images," In *Proc. Of the Conference Vision Interface 2000*, pages: 61-68, 2000.
- [4]. Can, H. Shen, J. N. Turner, H. L. Tanenbaum, and B. Roysam. "Rapid automated tracing and feature extraction from retinal fundus images using direct exploratory algorithms," In *IEEE Trans on Information Technology in Biomedicine*, pages: 125-138, 1999.
- [5]. Y. A. Tolia and S. M. Panas. "A fuzzy vessel tracking algorithm for retinal images based on fuzzy clustering," In *IEEE Trans on Medical Imaging*, 17, pages: 263-273, April 1998.
- [6]. Dorigo M, Maniezzo V. Alberto Colomi. "Ant system: Optimization by a colony of cooperating agents," In *IEEE Trans Systems, Man, and Cybernetics*, pages: 29- 41, 1996.
- [7]. T. A. Runkler, "Ant colony optimization of clustering models," In *International Journal of Intelligent Systems*, 20, pages: 1233-1251, 2005.
- [8]. R.J. Mullen, D. Monekosso, S. Barman, P. Remagnino, "A review of ant algorithms," In *Expert Systems with Applications*, pages: 9608-9617, 2009.
- [9]. C. Lorenz, J. Troccaz, E. Grimson, and R. M'osges, "Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2D and 3D medical images," In *Proc. CVRMed-MRCAS'97*, LNCS, pages: 233-242, 1997.
- [10]. Y. Sato, J. Troccaz, E. Grimson, and R. M'osges, "3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," In *Proc. CVRMed- MRCAS'97*, LNCS, pages 213-222, 1997.
- [11]. F. Mendels, C. Heneghan, J. P. Thiran, "Identification of the optic disk boundary in retinal images using active contours," *Proc. IMVIP*, pages: 103-115, 1999.
- [12]. P. C. Siddalingaswamy, G. K. Prabhu. "Automated Detection of Anatomical Structures in Retinal Images," In *Proc. ICCIMA*, pages: 164-168, 2007.
- [13]. Y. Sun, "Automated identification of vessel contours in coronary arteriogram by an adaptive tracking algorithm," In *IEEE Transactions on Medical Imaging*, pages: 78-88, 1989.

Estimating heart movement and morphological changes during robot-assisted coronary artery bypass graft interventions

Mathew A. Carias^{ab1}, Cristian A. Linte^{ab}, Daniel S. Cho^{ab}, and Terry M. Peters^{abc}

^aImaging Research Laboratories, Robarts Research Institute, London, Canada

^bSchulich School of Medicine and Dentistry, University of Western Ontario, London, Canada

^cCanadian Surgical Technologies and Advanced Robotics, London, Canada

ABSTRACT

Robot-assisted coronary artery bypass graft interventions rely on the assumption that pre-operatively acquired images and generated models represent the intra-procedure environment. This assumption can be misleading since the heart is composed of soft tissue that undergoes changes during the peri-operative workflow. Quantifying those changes during the peri-operative workflow can be of values during off-pump cardiac interventions since it allows us to track the movement of surgical targets intra-operatively to better predict their new location based on the pre-operative data. Here we present a method to quantify these changes from a global heart position perspective and a morphological feature change perspective. We use ultrasound images to identify the aortic and mitral valve annuli and measure their movement between different stages in the procedure. Based on these results, we can estimate the differences between the pre- and intra-operative anatomical features, how they may affect the position of surgical ports, and also identify the need to update or optimize the registration throughout the procedure workflow. We found that there are significant changes between all peri-operative stages that would affect the localization of surgical targets.

Keywords: Off-pump intra-cardiac interventions, Heart movement, Morphological changes, Valve annuli, Principal component analysis, Eigenvalue decomposition.

1 INTRODUCTION

During robot-assisted artery bypass graft procedures clinicians obtain pre-operative computed tomography (CT) scans to identify port locations that will optimize access to the target vessels with the use of robotic instruments [1]. These target vessels are not easily identified intra-operatively, which forces surgeons to essentially predict the location of the surgical targets using the pre-operative images. The peri-operative surgical workflow presented with this procedure may alter the position and morphology of the heart, which can cause surgical targets and features used in image-to-patient registration to alter in position. Additionally, information about feature and target locations must be updated intra-operatively to allow adequate alignment of surgical targets [2]. Measuring the overall heart movement and morphological changes of selected features during the peri-operative workflow is crucial for intra-operative planning and guidance to avoid the need of conversion to a traditional, open chest procedure [2].

This work will focus on estimating the anatomical changes that occur between the three stages during the peri-operative workflow of robot-assisted coronary artery bypass procedures. We will estimate these changes from a global heart positioning perspective and a selected-feature perspective (i.e. the mitral and aortic valves). Here we present an analysis of the changes in location of the heart and the variations in morphology of the features assisting with the model-to-subject registration. This information is not only critical to assess how different the pre-operative anatomy is relative to the intra-operative one, but also outlines the necessity to update and optimize the registration throughout the workflow for better results [3].

¹ Mathew Carias. E-mail: mcarias@uwo.ca, Telephone: +1(519) 663-5777.

2 METHODS

To measure the effects induced in the heart morphology during the peri-operative work-flow, the mitral and aortic valve annuli of patients undergoing robot-assisted coronary artery bypass were reconstructed using real-time ultrasound (US) imaging at three stages during the procedure: stage 1 – anesthetized and double lung ventilation; stage 2 – single lung ventilation; and stage 3 – 10 cm H₂O chest wall insufflations [4].

2.1 Peri-Operative Image Acquisition

Ultrasound images were acquired using a magnetically tracked trans-esophageal echocardiography (TEE) transducer (Agilent Technologies, Canada) and a 6 degree of freedom NDI Aurora (Northern Digital Inc., Canada) magnetic sensor coil embedded with the transducer, facilitating spatial tracking of the transducer with use of a magnetic field generator placed underneath the operating table [3].

Two-dimensional ultrasound images of the mitral and aortic valves were acquired at the same time point during the cardiac cycle - at mid-diastole. The imaging fan was rotated in 20 degree increments and a series of images were acquired for each valvular structure at each stage of the peri-operative workflow. The 2D images were then reviewed by an experienced echocardiographer and the mitral and aortic annuli were segmented by selecting corresponding points in each of the 2D images (**Figure 1**). Since the TEE transducer was spatially tracked, the acquired images and selected points corresponding to each segmented feature were positioned according to their spatial stamp within a common 3D coordinate system [5]. Ultimately the mitral and aortic valve annuli were reconstructed by connecting the selected points with a 3D spline, leading to a pair of annuli at each of the three work-flow stages, for each of the four patients who have undergone the procedure to date.

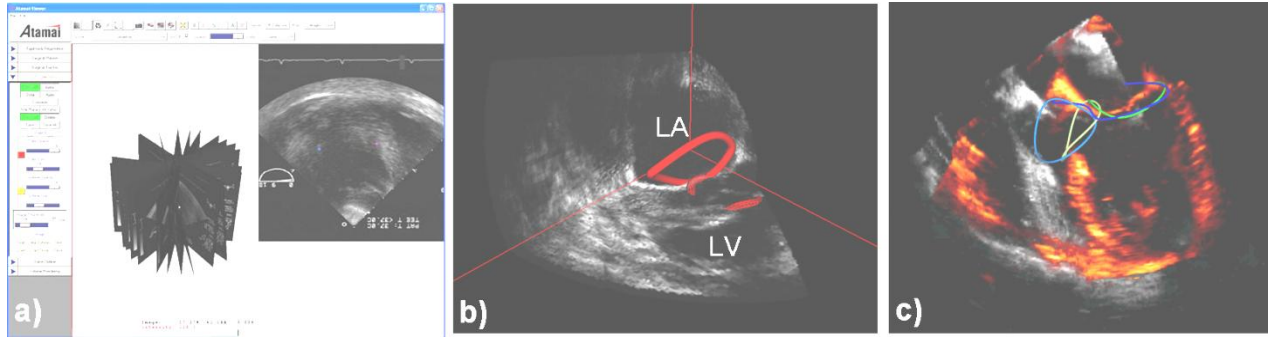


Figure 1. a) Interactive software tool used for US image collection, feature identification and anulus point collection; b) US image showing segmented mitral and aortic annuli; c) US images depicting the heart at two different stages in the workflow and showing the segmented mitral and aortic valve annuli.

2.2 Estimating Global Position and Morphological Changes

Once the data points were acquired, they were reordered and shaped to outline either the aortic or mitral valve. The positions of these points were reported and the centroid of each valve during each stage was computed by finding the mean of each of the individual x, y, and z component (Equation (1)). By tracing the movement of the centroid we are able to track the global motion of the heart via the location change of the anatomical features. We chose the centroid location as a tracking measure of the feature of interest, as it provides both the magnitude and direction of the displacement (Equation (2)).

$$Centroid_{x,y,z} = \left\langle \frac{1}{n} \sum_{i=1}^n x_i \mid \frac{1}{n} \sum_{i=1}^n y_i \mid \frac{1}{n} \sum_{i=1}^n z_i \right\rangle \quad (1)$$

$$Displacement = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \quad (2)$$

Morphological changes were explored via principal component analysis, singular value decomposition and eigenvalue decomposition. The principal component analysis procedure was performed on the centered annuli data, obtained by “translating” each annulus dataset to its standard position, where the translation vector was the negative of its corresponding centroidal position vector. A covariance matrix was then calculated for the centered set of data points and the geometric principal directions of the dataset were identified by performing an eigenvalue

decomposition of the covariance matrix, (Eqs. (3) and (4)). Here we have A being a linear transformation and x is the eigenvector and λ being the scalar value of the eigenvalue.

$$\text{covariance}(x|y|z) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})(z_i - \bar{z})}{n-1} \quad (3)$$

$$Ax = \lambda_x x \quad (4)$$

The eigenvectors corresponding to the largest and second largest eigenvalues represented the major and minor in-plane orientations of the annulus, respectively, while the eigenvector corresponding to the smallest eigenvalue represented the unit normal vector describing its out-of-plane characteristics [6]. An orthogonal plane of best fit was then computed for each annulus, using the eigenvector corresponding to the lowest eigenvalue as its normal vector describing its orientation. We then computed the relative angular orientation of the two valvular structures, using the dot product of the normal unit vectors corresponding to the two valvular planes of best fit, n_1 and n_2 , (Equation (5)).

$$\vec{n}_1 \cdot \vec{n}_2 = |\vec{n}_1| |\vec{n}_2| \cos \theta \quad (5)$$

3 RESULTS

3.1 Morphological Feature Characterization

We quantified any morphological changes by the use of principal component analysis, where the computed eigenvalues were associated with the effective length of the major and minor axes of each feature, as reported in **Table 1**. We did not report the out-of-plane axis since in many cases the associated eigenvalue was close enough to zero, showing that the feature did not present significant “out-of-plane” characteristics.

Table 1. Morphological information (Mean \pm Std. Dev.) of the mitral and aortic valve annulus at each stage of the peri-operative workflow of RA-CABG interventions. N=4 patients.

Workflow Stage	Mitral Valve Annulus			Aortic Valve Annulus			Inter Annular	
	Effective Length (mm)	Effective Major Axis(mm)	Effective Minor Axis(mm)	Effective Length (mm)	Effective Major Axis(mm)	Effective Minor Axis(mm)	Distance (mm)	Angle (deg)
Dual-lung ventilation	134 \pm 12.8	17.0 \pm 5.7	10.3 \pm 4.7	97 \pm 14.1	13.0 \pm 7.7	8.95 \pm 5.5	31 \pm 4.2	49 \pm 5.6
Single lung ventilation	123 \pm 13.0	16.5 \pm 7.1	11.1 \pm 3.9	83 \pm 8.9	10.8 \pm 6.3	8.12 \pm 2.2	33 \pm 4.8	46 \pm 4.7
Chest wall insufflation	123 \pm 4.7	15.1 \pm 7.3	10.7 \pm 5.1	110 \pm 11.3	14.5 \pm 7.5	9.50 \pm 6.9	31 \pm 4.6	75 \pm 11.4*

The morphological characterization of the mitral and aortic valve annuli revealed small variations in the effective perimeter and effective length of the major and minor axes of the valvular structures between the three stages for both types of procedure under investigation. The morphological variations of these features are important, as these valve annuli are used in the feature-based registration algorithm used to identify the transforms between subsequent stages in the procedure. If the geometry of the features changes drastically, then a rigid-body registration that only characterized the global movement of the heart may not be sufficient.

According to our measurements, consistent variations were observed throughout all datasets at Stage 1 and Stage 2, while larger variations were observed at Stage 3*. These variations resulted due to a poor definition of the valves in two of the patients, whose heart rates were too irregular to enable precise gating and acquisition of the US images at mid-diastole, resulting in an abnormally-reconstructed valvular geometry at that stage. However, considering that the images were reviewed off-line, after the procedure, the acquisition could not be repeated.

3.2 Overall Heart Movement

We then examined the overall heart movement on each valve individually. **Table 2** shows the overall heart changes of the mitral and aortic valve between specified stages. We specified the overall heart movement by the

centroid movement, angular normal change. The movement of the heart during the robot-assisted CABG procedure from stage 1 (anesthetized, dual-lung ventilation) to stage 2 (single-lung ventilation) averaged to 24.0 ± 3.0 mm for the mitral valve and 32.7 ± 9.6 mm for the aortic; from stage 1 to stage 3 (chest wall insufflation), the mitral valve experienced a movement of 29.5 ± 21.1 mm and the aortic movement averaged to 34.9 ± 25.7 mm.

Table 2. Mitral and aortic valve movement between workflow stages of robot-assisted coronary artery bypass graft interventions with \pm standard deviations.

Workflow Stage	Mitral Valve		Aortic Valve	
	Centroid Movement (mm)	Angular Change of Normal (deg)	Centroid Movement (mm)	Angular Change of Normal (deg)
Stage 1-2	24.0 ± 3.0	40.3 ± 25.2	32.7 ± 9.6	36.6 ± 9.6
Stage 2-3	40.0 ± 10.3	44.7 ± 24.8	39.4 ± 8.6	32.3 ± 8.6
Stage 1-3	29.5 ± 21.1	28.0 ± 6.3	34.9 ± 25.7	47.9 ± 25.7

In addition, we also quantified the change in orientation of the valvular plane based on the Euler angles computed using the orthonormal bases corresponding to each annulus at each stage in the procedure workflow. This change in orientation was reported in terms of the normal vector corresponding to the orthogonal plane of best fit for each valvular structure (**Table 2**). The direction of the corresponding normal vector and centroid location of each annulus are used in the registration algorithm used to identify the transformations between successive stages.

For a better interpretation of the global heart displacement observed in this study, we used a registration algorithm driven by the mitral and aortic valve annuli corresponding to different stages in the workflow to transform a model of a human heart obtained by segmenting a patient’s pre-operative CT scan. The stage-to-stage displacement transforms were identified by registering the datasets corresponding to subsequent stages using a feature-based registration previously developed and tested in the laboratory; these transforms were then applied to the cardiac model to update its position from one stage to another. **Figure 2** shows the model of the heart at Stage 1 (dual-lung ventilation) in red, along with its position and orientation at Stage 2 (single-lung ventilation) in green, and ultimately at Stage 3 (following chest insufflation) in blue.

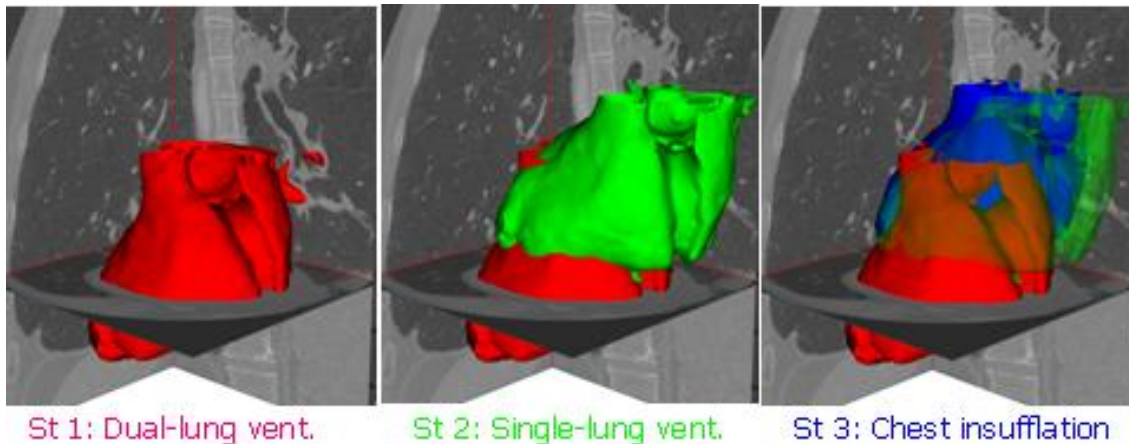


Figure 2. Visual representation showing an automatically segmented epicardial model of a patient’s heart animated using the sequential peri-operative transforms based on the valvular structures. Stage 1 is shown in the left panel, Stage 1 & 2 are shown in the middle panel and Stage 1, 2 & 3 are shown in the right panel. Note a latero-posterior displacement of the heart following lung deflation, followed by a posterior displacement of the heart after chest insufflations.

4 DISCUSSION AND CONCLUSION

In the context of the robot-assisted CABG procedures, we have currently analyzed the data from 4 patients from a total of over 60 patients who have agreed to participate in the study. Similarly, we have noticed substantial movement of the heart and valvular structures induced during the deflation of one lung and the insufflation of the chest. The overall heart displacement was on the order of 24 ± 3.0 mm for the mitral valve and 37.2 ± 9.6 mm for

the aortic valve from stage 1 to stage 2. We see similar movements between stages 1 and 3 where we saw an average movement of 29.5 ± 21.1 mm for the mitral valve and 34.9 ± 25.7 mm for the aortic valve. Global heart displacements on this order of magnitude are significant and hence the pre-operative plan needs to be updated to reflect these changes provide the clinician with the intra-operative location of the surgical targets.

The morphological characterization of the mitral and aortic valve annuli has also revealed small variations in the effective perimeter and effective length of the major and minor axes of the valvular structures between the three stages. Using the GraphPad Prism 4 statistical analysis package, we have performed a statistical comparison using a two-way analysis of variance (ANOVA) between the effective perimeter and effective long and short axes of the mitral and aortic annuli across the patient sample. The result have shown that no significant differences ($p > 0.05$) existed between these parameters at different procedure stage, except for those due to the patient variability, such as the size of the patients and their organs. Moreover, no significant differences were observed in the inter-annular distance ($p > 0.1$). These two observations together suggest that no significant morphological changes have occurred during the procedure workflow and therefore a rigid body registration may be sufficient to update the pre-operative surgical plan for robot-assisted CABG procedures.

The estimation of the global migration of the heart during the typical peri-operative workflow associated the robot-assisted CABG procedures allows us to identify the position and orientation of the patient's heart at the stage prior to therapy delivery and use this information to update an initial pre-operative plan derived solely based on a pre-operative dataset. In turn, this information can be used to update the intra-operative location of the target vessel – typically the LAD coronary artery, and moreover, the optimal port placement location to ensure that the minimally invasive procedure will not need conversion to traditional open-chest surgery. In addition, similar technique can be employed to study the hear migration patterns during other minimally invasive procedures, such as mitral valve replacement or atrial septal defect repair procedures. In fact, we are currently investigating the development of a registration algorithm that can make use of the peri-operative heart displacement and adequately predict the intra-operative location of surgical targets based on their pre-operatively determined location.

Acknowledgements

The authors would like to thanks Dr. Bob Kiaii and Dr. Dan Bainbridge for their help with clinical data acquisition and feature segmentation and John Moore and Chris Wedlake for their technical support. Also, we would like to acknowledge funding for this work provided by the Natural Sciences and Engineering Research Council, the Canadian Institutes of Health Research and the Heart and Stroke Foundation of Canada.

References

- [1] Martens, T. P., M.M. Hefti, R. Kalimi, C.R. Smith and M. Argenziano. "Robot-assisted off-pump minimally invasive reoperative coronary artery bypass grafting". Case report. *Heart Surg. Forum.* **7**, E533-4. 2004
- [2] Mierdl, S., C. Byhahn, V. Lischke, T. Aybek, G. Wimmer-Greinecker, G. Matheis, P. Kessler and K. Westphal 2002. Echocardiographic findings in minimally invasive coronary artery bypass grafting: The role of intrathoracic CO₂ - insufflation and single lung ventilation. *Heart Surg. Forum.* **5** Suppl 4, S398-419.
- [3] Linte, C. A., J. Moore, C. Wedlake, D. Bainbridge, G.M. Guiraudon, D.L. Jones and T.M. Peters. "Inside the beating heart: An in vivo feasibility study on fusing pre- and intra-operative imaging for minimally invasive therapy". *Int. J. Comput. Assist. Radiol. Surg.* **4**, 113-123, 2009.
- [4] Chiu, A. M., D. Dey, M. Drangova, W.D. Boyd and T.M. Peters. "3-D image guidance for minimally invasive robotic coronary artery bypass." *Heart Surg. Forum.* **3**, 224-231. 2000.
- [5] Linte, C. A., M. Wierzbicki, J. Moore, G. Guiraudon, D.L. Jones and T.M. Peters. "On enhancing planning and navigation of beating-heart mitral valve surgery using pre-operative cardiac models." *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2007, 475-478, 2007.
- [6] Vidal, R., Y. Ma and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1945-1959, 2005.

SPiD: An SVM-Based Protein Discriminator for Outer Membrane Proteins

Babak Alipanahi*

David R. Cheriton School of Computer Science, University of Waterloo
200 University Ave W., Waterloo, ON N2L 3G1, Canada

ABSTRACT

Discrimination of Outer Membrane Proteins (OMP) from other types of membrane and globular proteins is an important step in their secondary and tertiary structure prediction. Moreover, a reliable discrimination method can be used for whole genome analysis and hence discovery of new OMPs. In this paper, we propose an SVM-based protein discriminator for OMPs (SPiD) from other types of proteins, i.e., globular and inner membrane proteins. This approach uses amino acid and amino acid pair composition values, the length of protein sequence, and a newly defined feature called β -barrel score. When applied to a dataset consisting of 1,087 proteins, SPiD achieves an overall accuracy of 96%; to the best of authors' knowledge, this is higher than the accuracy of other previous studies. When SPiD is trained to pick up only outer membrane β -barrels, it reaches an overall accuracy of 99%.

Keywords: SVM, protein classification, membrane proteins

1 INTRODUCTION

The most remarkable fact about Gram-negative bacteria is their cell envelopes. It consists of two layers: Inner Membrane (IM) and Outer Membrane (OM), which are separated by periplasm. IM is in direct contact with cytoplasm and periplasm while OM is in contact with extracellular environment and periplasm [1]. Integral IM proteins span the membrane by α -helices while integral OM proteins span the membrane by β -strands and form a β -barrel. OM Proteins (OMPs) have diverse functions and are divided into several families: selective active and passive transporters of molecules, enzymes, defense proteins, structural proteins, and toxins.

Several methods for discrimination of OMPs have been proposed in the recent years. These methods can be divided into three major groups: methods using sequence alignment information and/or HMM [2, 3, 4, 5, 6], methods based on amino acid composition values [7, 8, 9], and methods using amino acid sequence properties like estimated folding pseudo-energy or average hydrophobicity [10, 11, 12]. In this paper, we propose an SVM-based protein discriminator for OMPs (SPiD), using amino acid and amino acid pair composition values, sequence length, and a feature especially tailored for β -barrels.

2 Materials and methods

2.1 Datasets

The dataset used here is the same as in [9], since it is one of the most challenging and comprehensive available datasets. Moreover, the authors report the best results to date and it facilitates the fair comparison of our results with those of previous studies. This dataset primarily consists of 377 OMPs and 268 α -helical membrane proteins extracted from PSORT-B database [13], and 674 globular proteins from the PDB40D_1.37 database of SCOP [14, 15]. It is filtered for sequence identity of less than 40% using CD-HIT algorithm [16]. The resulting dataset has 208 OMPs, 206 α -helical membrane proteins, and 673 globular proteins consisting of all- α , all- β , $\alpha + \beta$, and α/β proteins. Herein after, we will define the following notation for the dataset used for discrimination: the set of outer membrane proteins (OMP), the set of α -helical membrane proteins (TMH), the set of globular proteins (GLB), and the set of non-OMPs (NOM).

In order to verify the capability of the proposed approach in discrimination of transmembrane β -barrels from α -helical membrane and other non- β -barrel proteins, we also used `TMFDB_alpha_non_redundant` and `TMFDB_beta_non_redundant` datasets [17]. These datasets consist of proteins with experimentally known structures that are filtered

*Corresponding author e-mail: balipana@cs.uwaterloo.ca

for sequence similarity of less than 30%, using CLUSTALW version 1.81 [18]. TMPDB_alpha_non_redundant contains 231 and TMPDB_beta_non_redundant contains 15 proteins.

2.2 Features

In a protein sequence of length N , for amino acids a and b , the peptide (amino acid) and dipeptide (amino acid pair) composition values are defined by

$$C_a = \frac{n_a}{N}, \quad D_{ab} = \frac{p_{ab}}{N-1}, \quad (1)$$

where n_a and p_{ab} are the number of occurrences of amino acid a and amino acid pair ab in the sequence, respectively. OMPs usually have longer amino acid sequences than some of globular proteins, since for example forming a β -barrel structure requires a minimum number of amino acids, so it was added too. We denote peptide composition value features by C, dipeptide features by D, and sequence length feature by L. There are 20 ‘‘C’’ and 400 ‘‘D’’ features.

We define a new feature called β -barrel score whose original idea is taken from [4]. Every amino acid in the membrane spanning section of the protein sequence can be either Lipid Exposed (LE) or barrel Interior Exposed (IE). The β -strand score for every position i , B_i , is defined as [4]:

$$B_i^1 = \sum_{j \in \mathcal{E}} L(a_{i+j} | a_{i+j} \in \text{IE}) + \sum_{j \in \mathcal{O}} L(a_{i+j} | a_{i+j} \in \text{LE}), \quad (2)$$

$$B_i^2 = \sum_{j \in \mathcal{E}} L(a_{i+j} | a_{i+j} \in \text{LE}) + \sum_{j \in \mathcal{O}} L(a_{i+j} | a_{i+j} \in \text{IE}), \quad (3)$$

and $B_i = \max(B_i^1, B_i^2)$; where in equations (2) and (3), $L(a_{i+j} | a_{i+j} \in \text{IE}) = \log(\text{Pr}\{a_{i+j} | a_{i+j} \in \text{IE}\})$, which is estimated from the real data; the same is true for the LE case. Moreover, the set of even and odd shifts are defined as $\mathcal{E} = \{0, 2, 4, 6, 8\}$ and $\mathcal{O} = \{1, 3, 5, 7, 9\}$. When summing $L(a_{i+j} | a_{i+j})$ values, we actually multiply the corresponding probabilities. If we assume independence between consequent residues, B_i^k , $k = 1, 2$ values are the logarithm values of the probability that a β -strand starts at residue i with different assumptions, whether it is IE or LE. It turns out that a window size of ten is optimum. After calculating B_i for all residues 1 to $N - 9$; we form the new feature called β -barrel score (B) which is defined as $B = \frac{1}{N-9} \sum_{i=1}^{N-9} B_i^2$.

In Table 1, mean values of 20 peptide composition values, sequence length, and β barrel score features are listed for GLB, TMH, NOM, and OMP datasets. To perform feature selection, we used backward elimination-forward selection (BE-FS) method. Since running backward elimination on 400 dipeptide features is not feasible, we only ran forward selection on them.

3 Results and Discussion

3.1 Implementation

We have used the LIBSVM software package¹, which is both very fast and reliable. In each scenario, we repeated each experiment for 100 times and each time randomly changed the permutation of proteins. The standard deviation of the calculated results is approximately 0.1%. In order to optimize the RBF kernel parameters, we have used grid optimization technique in two coarse and fine steps. It turned out that optimal values for penalty parameter and kernel width were 10 and 2, respectively.

3.2 Performance evaluation

In order to participate all data points in the performance evaluation process, we use 5-fold cross validation. In order to compare SPiD’s results with previous studies, we use measures that have been widely used before. Suppose TP, FP, TN, and FN denote true positive, false positive, true negative and false negative assignments, respectively. By *positive* we mean a correctly classified protein, and by *negative* we mean an incorrectly classified protein. We use the following widely-used performance measures:

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{SPC} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \quad (4)$$

where SEN, SPC and ACC stand for sensitivity (ability to discover OMPs, a small sensitivity value indicates that many OMPs will not be discovered), specificity (ability to correctly sift OMPs from non-OMPs), and overall accuracy which is an indicator of overall performance. Moreover, we use the Matthew’s correlation coefficient (MCC) which is a better performance measure defined as follows [19]:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (5)$$

MCC value is zero for a completely random assignment and one for a perfect discrimination.

¹Available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

Table 1. Mean values of features

Feature	GLB	TMH	NOM	OMP
L	183	424	240	552
B	0.82	1.06	0.88	1.32
Ala	8.42	10.27	8.86	9.37
Arg	5.05	4.45	4.91	5.23
Asn	4.40	3.06	4.09	5.44
Asp	5.84	3.32	5.25	5.88
Cys	1.47	0.85	1.32	0.41
Gln	3.94	3.21	3.77	4.71
Glu	6.72	3.74	6.02	4.86
Gly	7.65	8.33	7.81	8.69
His	2.20	1.68	2.08	1.25
Ile	5.77	7.50	6.17	4.72
Leu	8.50	12.72	9.49	8.94
Lys	6.17	3.38	5.52	4.89
Met	2.19	3.58	2.52	1.66
Phe	3.77	5.49	4.17	3.75
Pro	4.47	4.29	4.43	3.74
Ser	5.77	5.88	5.79	8.04
Thr	5.71	5.20	5.59	6.31
Trp	1.34	2.05	1.51	1.24
Tyr	3.42	2.82	3.28	4.13
Val	7.20	8.19	7.43	6.75

All composition values are in percentile.

Table 2. OMP-GLB discrimination results

Features	SEN	SPC	ACC	MCC
L+20C+B	89.3	98.4	96.1	0.896
L+9C+B	89.0	98.9	96.5	0.904
L+9C+2D+B	89.9	98.8	96.6	0.908

Table 3. OMP-TMH discrimination results

Features	SEN	SPC	ACC	MCC
L+20C+B	95.9	93.2	94.6	0.892
L+10C	96.3	94.7	95.5	0.910
L+10C+3D	98.2	96.3	97.3	0.945

Table 4. OMP-NOM discrimination results

Features	SEN	SPC	ACC	MCC
L+20C+B	86.2	97.7	95.4	0.855
L+13C+B	85.0	98.3	95.6	0.861
L+13C+3D+B	87.0	98.2	96.0	0.872

SEN, SPC, and ACC are in percentile.

The best results are shown in bold.

First line, shows the performance measures without dipeptide features; second line after backward elimination; and the third line after forward selection ran only on dipeptide composition features.

3.3 Discrimination of OMPs

We experimented discrimination of OMPs in three scenarios: from globular proteins (OMP-GLB), from α -helical membrane proteins (OMP-TMH), and from non-OMPs (OMP-NOM).

3.3.1 OMP-GLB and OMP-TMH discrimination

In Table 2, performance results of OMP-GLB discrimination are listed. Initial ACC and MCC values are better than previous studies' results; while backward elimination and forward selection even more improved these results. After backward elimination, remaining amino acids were aromatic (Trp and Tyr) and polar residues (Cys, Gln, Pro, His and Thr). The added dipeptide composition features were Asp-Phe and Tyr-Asn, both a combination of a polar and an aromatic residue, which are abundant in OMPs.

In OMP-TMH scenario (results listed in Table 3), backward elimination improved ACC and MCC by 0.9% and 0.018, respectively; and removed β -barrel score and half of the peptide composition values. The remaining residues were good α -helix formers (Ala and Met), bad α -helix formers (Gly, Pro, Ser, Tyr), and Asp, His, Ile and Lys. Forward selection added Gln-Ala (polar-aliphatic), Asp-Ala (charged-aliphatic), and Glu-Phe (charged-aromatic), and boosted ACC and MCC values by 1.8% and 0.035, respectively.

3.3.2 OMP-NOM discrimination

OMP-NOM discrimination is more important than other scenarios, therefore, we elaborate more on it. Discrimination accuracy using all 20 amino acid composition values, length of protein sequence, and β -barrel score was quite acceptable and better than all previous studies. Backward elimination, did not improve the performance so much but reduced the number of features from 22 to 15; most of the omitted features had close mean values. Forward selection, added 3 dipeptide features: Asp-Ala (charged-aliphatic), Glu-Phe (charged-aromatic) and Asn-Lys (polar-charged); These added features boosted ACC and MCC values from 95.6% and 0.861 to 96% and 0.872, respectively. It is important that sensitivity was improved from 85% to 87%. Detailed performance measures are listed in Table 4. The relation between MCC and SEN values for different values of SPC for OMP-NOM discrimination is depicted in Figure 1. It can be seen that MCC value nearly grows linearly with the growth of SEN value.

In order to better analyze the performance, for each protein, the probability of being classified incorrectly is estimated by running the discrimination experiment 500 times and counting the number of times that any protein is classified incorrectly, whenever it is in the validation dataset. The estimated probabilities are depicted in Figure

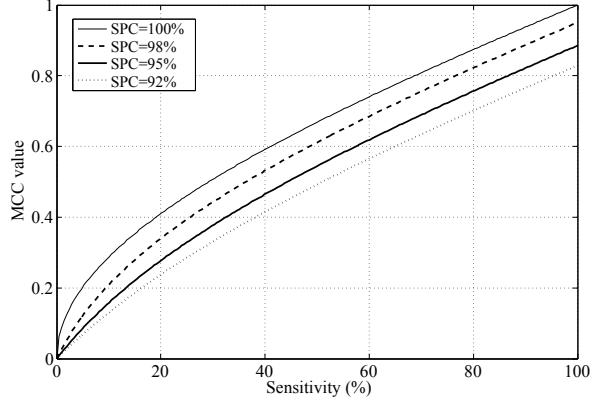


Figure 1. MCC values vs. sensitivity for different values of SPC for OMP-NOM discrimination.

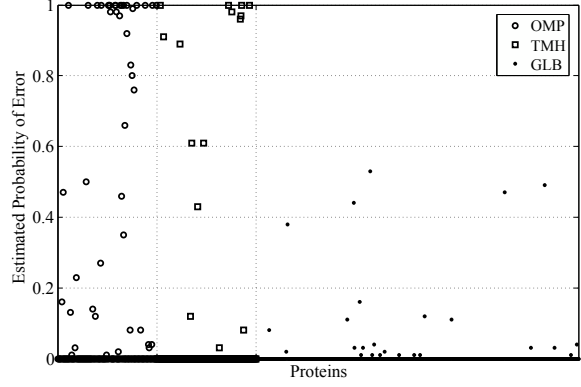


Figure 2. Estimated probability of error for all proteins.

2. It is interesting that some proteins (mostly OMPs) are always misclassified, i.e., their estimated probability of error is one and some proteins (mostly globular proteins) are never misclassified. The detailed results of estimated misclassification error probabilities are listed in Table 7.

3.4 Discrimination of β -barrels

In a more specific discrimination scenario, we apply our approach to discrimination of transmembrane β -barrels from other types of proteins. To do so, we have used the TMPDB.alpha.non_redundant, TMPDB.beta.non_redundant and GLB datasets. It is interesting that the mean β -barrel score feature of β -barrels (1.79) is nearly twice as other datasets (0.81 for GLB, 0.98 for TMH, and 0.85 for α -helical membrane and globular proteins dataset), and has a very large mean difference in all scenarios. It is mainly because this feature is especially tailored for β -barrels. We use the same set of features that are found after performing backward elimination-forward selection in OMP-NOM discrimination. Performance results are listed in Table 6. In all cases discrimination accuracy is very high.

Table 5. Comparison with other studies.

Method	Scenario	SEN	SPC	ACC
SPiD (SVM)	OMP-GLB	89.9	98.8	96.6
	OMP-TMH	98.2	96.3	97.3
	OMP-NOM	87.0	98.2	96.0
[7] (Linear Classifier)	OMP-GLB	85.5	92.5	92.1
[9] (SVM)	OMP-GLB	88.0	90.4	94.4
	OMP-TMH	99.0	92.7	95.9
	OMP-NOM	90.9	94.7	93.9
[12] (Neural Network)	OMP-GLB	83.7	97.6	94.3
	OMP-TMH	91.8	91.7	91.8
	OMP-NOM	81.3	97.5	94.4

Table 6. Results of β -barrel discrimination.

Scenario	SEN	SPC	ACC	MCC
BB-GLB	75.3	99.6	98.9	0.782
BB-AA	95.8	100	99.6	0.974
BB-NBB	80.0	99.8	99.3	0.829

BB: TMPDB.beta.non_redundant
AA: TMPDB.alpha.non_redundant
NBB: AA + GLB

Table 7. Estimated probability of error analysis.

Dataset	size	$P_e = 0$	$0 < P_e < 1$	$P_e = 1$
GLB	673	0.96	0.04	0.00
TMH	206	0.93	0.05	0.02
OMP	208	0.78	0.13	0.09
Total	1087	0.92	0.06	0.02

4 Discussion and Conclusion

In this study, we proposed SPiD, a new SVM-based approach for discrimination of OMPs and in a more specific case, transmembrane β -barrels. By adding two features to the peptide and dipeptide composition values and performing feature selection, SPiD was proved to be very accurate. Park *et al.* proposed a method based on amino acid and amino acid pair composition values and reported an accuracy of 93.9% in a set of 208 OMPs that was calculated using 5-fold cross validation [9]. Gromiha and Suwa developed a method based on amino acid properties and reported a prediction rate of 94.4% [12]. In comparison to the aforementioned studies, SPiD achieved OMP-GLB, and OMP-TMH, OMP-NOM discrimination accuracies of 96.6%, 97.3%, and 96.0%, respectively. When trained to pick only β -barrels, SPiD was able to reject 913 out of 919 non- β -barrels and cover 12 out of 15 β -barrel families. Moreover,

by performing error probability analysis, it turned out that some proteins were always misclassified because they were more similar to non-OMPs.

References

- [1] N. Ruiz, D. Kahne, and T. J. Silhavy, “Advances in understanding bacterial outer-membrane biogenesis,” *Nature reviews. Microbiology* **4**, pp. 57–66, January 2006.
- [2] T. V. Gnanasekaran, S. Peri, A. Arockiasamy, and S. Krishnaswamy, “Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins,” *Bioinformatics* **16**, pp. 839–842, September 2000.
- [3] P. L. L. Martelli, P. Fariselli, A. Krogh, and R. Casadio, “A sequence-profile-based hmm for predicting and discriminating beta barrel membrane proteins,” *Bioinformatics (Oxford, England)* **18 Suppl 1**, 2002.
- [4] W. C. Wimley, “Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures,” *Protein science : a publication of the Protein Society* **11**, pp. 301–312, February 2002.
- [5] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, “A hidden markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins,” *BMC bioinformatics* **5**, March 2004.
- [6] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, and B. Rost, “Predicting transmembrane beta-barrels in proteomes,” *Nucl. Acids Res.* **32**, pp. 2566–2577, May 2004.
- [7] Q. Liu, “Identification of β -barrel membrane proteins based on amino acid composition properties and predicted secondary structure, url = [http://dx.doi.org/10.1016/S1476-9271\(02\)00085-3](http://dx.doi.org/10.1016/S1476-9271(02)00085-3), volume = 27, year = 2003,” *Computational Biology and Chemistry*, pp. 355–361, July.
- [8] A. G. Garrow, A. Agnew, and D. R. Westhead, “Tmb-hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins,” *BMC bioinformatics* **6**, 2005.
- [9] K.-J. J. Park, M. M. Gromiha, P. Horton, and M. Suwa, “Discrimination of outer membrane proteins using support vector machines,” *Bioinformatics (Oxford, England)* **21**, pp. 4223–4229, December 2005.
- [10] Y. Zhai and M. H. Saier, “The beta-barrel finder (bbf) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes,” *Protein science : a publication of the Protein Society* **11**, pp. 2196–2207, September 2002.
- [11] J. Waldspühl, B. Berger, P. Clote, and J.-M. M. Steyaert, “Predicting transmembrane beta-barrels and inter-strand residue interactions from sequence,” *Proteins* **65**, pp. 61–74, October 2006.
- [12] M. M. Gromiha and M. Suwa, “Influence of amino acid properties for discriminating outer membrane proteins at better accuracy,” *Biochimica et biophysica acta* **1764**, pp. 1493–1497, September 2006.
- [13] J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. L. Brinkman, “Psort-b: improving protein subcellular localization prediction for gram-negative bacteria,” *Nucl. Acids Res.* **31**, pp. 3613–3617, July 2003.
- [14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “Scop: a structural classification of proteins database for the investigation of sequences and structures,” *Journal of molecular biology* **247**, pp. 536–540, April 1995.
- [15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The protein data bank,” *Nucleic acids research* **28**, pp. 235–242, January 2000.
- [16] W. Li, L. Jaroszewski, and A. Godzik, “Clustering of highly homologous sequences to reduce the size of large protein databases,” *Bioinformatics* **17**, pp. 282–283, March 2001.
- [17] M. Ikeda, M. Arai, T. Okuno, and T. Shimizu, “Tm pdb: a database of experimentally-characterized transmembrane topologies,” *Nucl. Acids Res.* **31**, pp. 406–409, January 2003.
- [18] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, “Multiple sequence alignment with the clustal series of programs,” *Nucleic acids research* **31**, pp. 3497–3500, July 2003.
- [19] B. W. Matthews, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme,” *Biochimica et biophysica acta* **405**, pp. 442–451, October 1975.

Application of a respiratory CT sequence's combined histogram to estimate intra-sequence lung's air volume variations

Ali Sadeghi Naini^{a, b, c}, Rajni V. Patel^{a, c, d}, and Abbas Samani^{a, b, e}

^aDepartment of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada

^bImaging Research Laboratories, Robarts Research Institute (RRI), London, ON, Canada

^cCanadian Surgical Technologies & Advanced Robotics (CSTAR), London, ON, Canada

^dDepartment of Surgery, The University of Western Ontario, London, ON, Canada

^eDepartment of Medical Biophysics, The University of Western Ontario, London, ON, Canada

ABSTRACT

A technique is proposed to segment the lung's air voxels in an image sequence based on a novel image sequence analysis. The concept involves using the image sequence's combined histogram to estimate the lung's air volume and its variations throughout respiratory CT image sequences. Accurate estimation of these parameters is very important in many applications related to lung disease diagnosis and treatment systems (*e.g.* brachytherapy) where the air volume and its variations are either the variables of interest themselves or are dependent/independent variables. *Ex vivo* experiments were conducted on porcine left lungs in order to demonstrate the performance of the proposed technique. The proposed method was validated using a breath-hold CT image sequence with known air volumes inside the lung. The results indicate a very good ability of the proposed method for estimating the lung's air volume and its variations in a respiratory image sequence.

Keywords: Air, Volume, Segmentation, Respiratory, CT, Sequence, Lung, Brachytherapy

1 INTRODUCTION

Estimation of lung air volume and/or its variations throughout a respiratory sequence has been proposed by several groups in several applications [1, 2]. However, what seems to be a major shortcoming in most of these studies is the lack of a more reliable non-empirical approach to obtain customized upper and lower segmentation threshold values. Such systematic approach can replace existing empirical approaches that usually hamper segmentation accuracy. Since empirical approaches for finding upper and lower threshold values for accurate lung's air segmentation are usually unavailable, threshold values are often set to segment both lung tissue and air (whole lung). The air volume variations in the sequence are then estimated by calculating the whole lung volume differences within the image sequence. In this approach, however, the lung's air volume in each image needs to be estimated from the whole lung volume, or its corrected version using another empirical correction factor. This usually results in higher errors in estimating both the lung's air volume and its variations throughout the sequence.

There are other applications; *e.g.* lung brachytherapy systems, where the lung's air volume and/or its variations during a respiratory sequence could be used as either a dependent [3, 4] or independent [5] variable. However, in some cases, due to lack of a reliable method to track the lung's air volume variations, one might prefer to use other variables which are correlated with the volume changes while being easier to measure or track accurately [6]. In a recent study conducted in our laboratory, Sadeghi Naini *et al* proposed a novel method to reconstruct the CT image of a totally deflated lung based on its partially inflated images [5]. Such a CT image would be very useful in performing tumor ablative procedures such as brachytherapy [7] for treating lung cancer. These procedures are usually performed after the target lung is completely deflated before starting the surgery. This implies that the lung physical domain would be no longer represented accurately by the pre-operative CT images. The proposed method consisted of acquiring a number of pre-operative breath-hold CT images at different lung volumes controlled by a ventilator and/or a volume-meter transducer. Each two successive CT images in the sequence were, then, registered with each other to obtain the registration parameters. Subsequently, each registration parameter was described as a function of the lung's air volume variation. Registration parameters corresponding to the totally deflated lung were then determined using extrapolation. Finally, the CT image of the totally deflated lung was reconstructed by registering the pre-operative image of the least inflated lung using the extrapolated parameters.

Although the concept proposed in this study was proven to be effective, its implementation using static breath-hold CT images may not be practical in clinical settings. In contrast to the static breath-hold imaging protocol, the free-breathing 4DCT is more suitable in the clinic as it is more straightforward to implement and less time consuming while being more convenient for patients. However, since there is no control mechanism over the lung's inhaled air volume while free-breathing 4DCT images are acquired, the lung's air volume corresponding to the image set is unknown. Hence, in order to apply the extrapolation technique in conjunction with the free-breathing 4DCT imaging protocol, an effective technique for estimating the lung's air volume and its variations is a paramount necessity.

In this paper, a technique for accurate image sequence segmentation is introduced based on a novel image sequence analysis. The concept is equally useful for segmenting image sequences, both static and dynamic. As described in Section 2, this concept is proposed to estimate the lung's air volume and its variations in respiratory CT image sequences using sequence combined histogram. *Ex vivo* experiments were conducted on porcine left lungs in order to demonstrate the validity of the proposed method. The proposed method was validated using a breath-hold CT image sequence with known lung's air volumes. The experiments conducted and the results obtained are presented in Section 3. As discussed and concluded in Section 4, the obtained results indicate a very favorable ability of the proposed technique for estimating the lung's air volume and its variations in a respiratory image sequence.

2 METHOD

2.1 Preliminaries

Image segmentation is defined as the process of assigning each image pixel/voxel to a specific class. Methods for performing image segmentation vary widely from simple techniques to complex algorithms. Typically, they depend on the specific application, imaging modality, and other factors. There is currently no general purpose segmentation method that yields acceptable results for any medical image. Although there are more general methods that can be applied for various types of images, methods that are specialized to particular applications can often achieve better performance by taking *a priori* knowledge into account. In many segmentation methods, finding proper algorithm parameters, *e.g.* threshold, initial seed, *etc.*, is a key step. However, these parameters are usually found empirically. This usually results in significant errors during the segmentation process.

Thresholding is a simple, yet often effective segmentation technique which divides the image into desired classes by comparing each image pixel/voxel value with a number of intensity values called thresholds. The most important step in the thresholding method is fine tuning the threshold values. These values have significant influence on the accuracy of the segmentation algorithm. As mentioned before, in many applications involving biomedical imaging procedure, this step is frequently implemented empirically. However, as suggested in the next section, a proper threshold value can be determined systematically for a given image sequence by simple analysis of the image sequence.

2.2 Air volume estimation algorithm

Fig. 1 shows a block diagram of the proposed method for estimating the lung's air volume and its variations in a respiratory CT image sequence. While simple, the main idea is quite effective. The concept takes advantage of the fact that the segmentation classes appear in all images in the sequence, though with variable shape and size. More specifically, each image in the sequence mainly consists of three segmentation classes including background, lung's air, and soft tissue. Considering mass conservation of the lung's air during respiration, a reduction in the volume of the background leads to equal increase in the volume of air in the lung and vice versa. In addition, taking into account the soft tissue incompressibility or near incompressibility, the volume of the soft tissue is (almost) constant throughout the image sequence. The proposed technique employs these constraints in order to find the best segmentation thresholds for a variable class throughout the sequence.

The algorithm starts with the input block where the whole image sequence is input. In the first step, histograms of all images within the sequence are calculated separately. These histograms are then passed to the second block where they are overlaid in order to form the sequence's combined histogram. After smoothing the histogram curves in order to remove high frequency noise-like variations, the convergence points in the sequence's combined histogram are extracted in the next block. The convergence points are defined as those intensity values within the combined histogram where all the separate histograms converge together, *i.e.* the distances between them tend to a minimum. These points are the optimum points to be used as the segmentation thresholds, since they best satisfy all the images' histograms besides satisfying the air mass conservation and tissue incompressibility constraints. After segmentation is performed using the determined threshold values in the fourth block, the fifth block counts the voxels segmented as the lung's air for each image separately. In the sixth block, the lung's air volume is calculated for each image by multiplying the number of voxels counted as the air by the voxel size. Finally, the

last block calculates the air volume variations within the sequence by subtracting the lung's air volumes between successive images. The experiments conducted to validate this approach are presented in the next section.

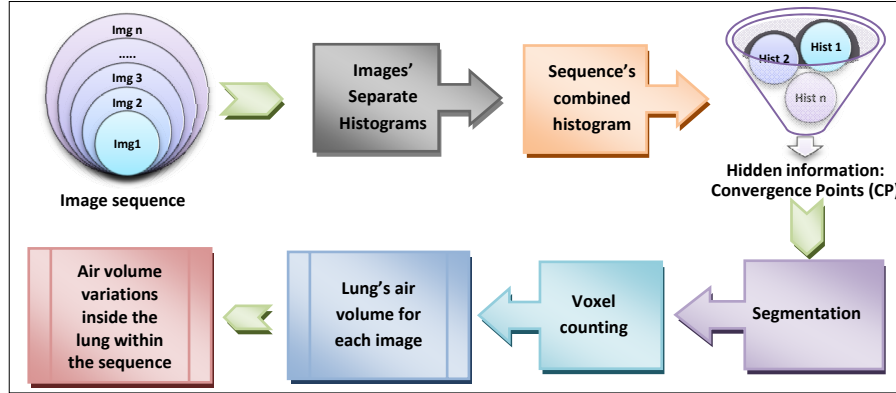


Figure 1. Block diagram of the algorithm proposed to estimate the lung's air volume and its variations in a respiratory CT image sequence.

3 Experiments and results

Ex vivo experiments were conducted on a porcine left lung in order to validate the proposed technique. The experiments were conducted using a number of static breath-hold CT images from a respiratory sequence acquired while the lung's air volume was controlled and known in each image. The lung obtained from an adult ~80 kg pig was inflated using an intra-trachea tube and a North American Drager Narkomed 2A ventilator machine. The air volume inside the lung was controlled by the ventilator. Micro-CT imaging was performed using a GE Locus Ultra scanner. The static breath-hold CT images of the lung were acquired at volumes of 700 ml, 600 ml, and 300 ml, respectively. The final images size was (228x186x324) voxels with a voxel size of (0.62³) mm³. These three 3D images were fed to the air volume estimation algorithm as the respiratory image sequence. Fig. 2 shows the combined sequence histogram obtained for these images. In this figure, the second hill belongs to the lung's air voxels in different images. The convergence points at the beginning and the end of this hill are indicated by arrows. As mentioned before, these points are the optimum points to be used as the upper and lower thresholds for segmenting the lung's air since they best satisfy all the images' histograms.

Fig. 3 demonstrates one middle slice of the CT images acquired at different volumes where the air inside the lung is segmented using the obtained threshold values. The lung's air volumes calculated based on the performed segmentation are given in Table 1. The table indicates that the estimation errors range from 5% to 6.3%, which is reasonably low. In other words, the accuracy of the proposed method for estimating the lung's air volume in a respiratory sequence is sufficiently good.

1 Discussion and conclusions

In this paper, a novel concept of image sequence analysis was introduced in order to obtain appropriate lower and upper threshold bounds for threshold-based lung image segmentation. This concept is equally useful for segmenting both static and dynamic image sequences. In this research, the concept was utilized to estimate the lung's air volume and its variations in respiratory CT image sequences using a combined sequence histogram. *Ex vivo* experiments were conducted on porcine left lungs in order to prove the concept. The proposed method was validated using a breath-hold CT image sequence with known lung air volumes. The obtained results indicated a very good ability of the method for estimating the lung's air volume and its variations throughout a respiratory image sequence. Considering its favorable capabilities, this technique can be used effectively in clinical applications such as lung brachytherapy where the lung's air volume and/or its variations in a respiratory sequence are needed.

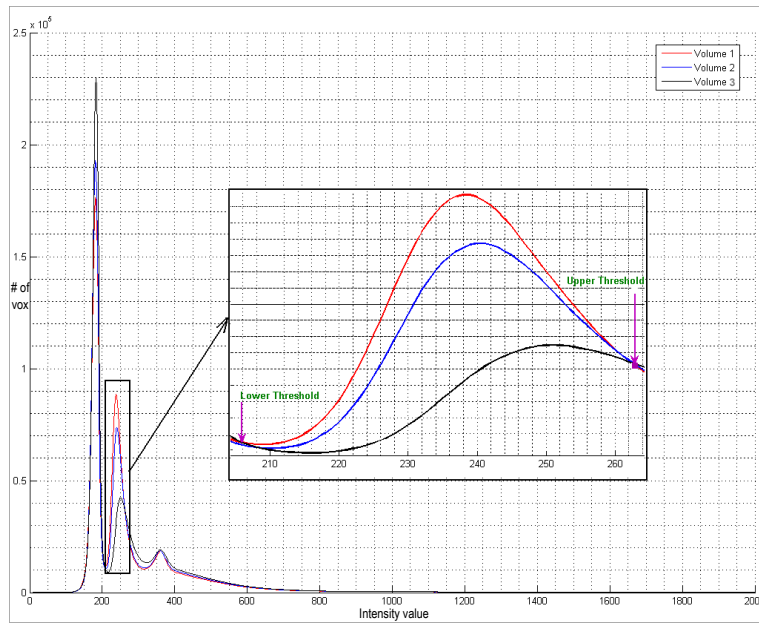


Figure 2. Combined sequence histogram for a respiratory sequence consisted of three static breath-hold CT images acquired at 700, 600, and 300 ml, respectively. The figure has been zoomed in to focus on the region of interest within the original combined histogram; lower and upper segmentation thresholds are indicated by arrows.

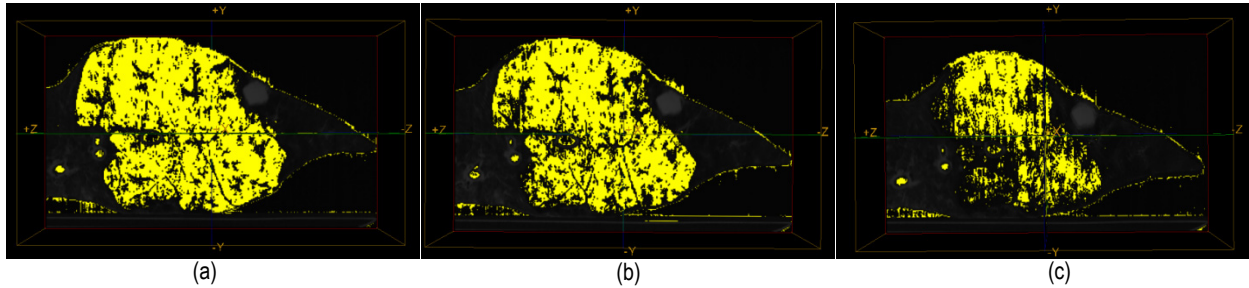


Figure 3. One middle slice of the static breath-hold CT images acquired at: (a) 700 ml, (b) 600 ml, (c) 300 ml; the air inside the lung is segmented using the lower and upper threshold values extracted from the sequence's combined histogram. The bright and dark regions show the air and soft tissue with the background, respectively.

Table 1. Summary results of estimated lung's air volumes in the respiratory CT sequence.

Image #	Air volume inside the lung	Estimated air volume inside the lung	Error
1	700 ml	665 ml	5%
2	600 ml	562 ml	6.3%
3	300 ml	319 ml	6.3%

AKNOWLEDGMENT

This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and by infrastructure grants from the Canada Foundation for Innovation awarded to the London Health Sciences Centre (Canadian Surgical Technologies & Advanced Robotics-CSTAR) and the University of Western Ontario.

REFERENCES

- [1] H.U. Kauczor, C.P. Heussel, B. Fisher, et al, "Assessment of lung volumes using helical CT at inspiration and expiration: comparison with pulmonary function tests", *AJR*, vol. 171, pp.1091-1095, 1998.
- [2] M.L. Goris, H.J. Zhu, F. Blankenberg, et al, "An automated approach to quantitative air trapping measurements in mild cystic fibrosis", *Chest*, vol. 123, pp. 1655-1663, 2003.
- [3] J.M. Reinhardt, K. Ding, K. Cao, et al, "Registration-based estimates of local lung tissue expansion compared to xenon CT measures of specific ventilation", *Med. Image Anal.*, vol. 12, pp. 752-763, 2008.
- [4] G. Li, N.C. Arora, H. Xie, et al, "Quantitative prediction of respiratory tidal volume based on the external torso volume change: a potential volumetric surrogate", *Phys. Med. Biol.*, vol. 54, pp. 1963-1978, 2009.
- [5] A. Sadeghi Naini, R. V. Patel, A. Samani, "CT image construction of the lung in a totally deflated mode", 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2009), Boston, Massachusetts, USA, pp. 578-581, 2009.
- [6] J.R. McClelland, J.M. Blackall, S. Tarte, et al, "A continuous 4D motion model from multiple respiratory cycles for use in lung radiotherapy", *Med. Phys.*, vol. 33, pp. 3348-58, 2006.
- [7] A.L. Trejos, A. W. Lin, M. P. Pytel, R. V. Patel, R. A. Malthaner, "Robot-assisted minimally invasive lung brachytherapy", *International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 3, pp. 41-51, 2007.

A fast breast nonlinear elastography reconstruction technique using the Veronda-Westman model

Mohammadhosein Amooshahi^a and Abbas Samani^{abc}

^aDepartment of Electrical & Computer Engineering, University of Western Ontario, London, ON, Canada;

^bDepartment of Medical Biophysics, University of Western Ontario, London, ON, Canada;

^cImaging Research Laboratories, Robarts Research Institute, London, ON, Canada

ABSTRACT

A common weakness of most conventional imaging modalities is that although they can detect the presence of pathological tissues, they are incapable of classifying tumors and determining whether they are malignant. To address this major issue, elastography has been developed. This is an imaging technique that provides the spatial distribution of tissue stiffness. The main idea behind elastography is the fact that tissue pathological changes such as those associated with cancer trigger significant changes in the tissue mechanical properties. The mechanical behavior of a tissue can be described by parameters characterizing its linear or nonlinear behavior. While soft tissues demonstrate linear behavior under small strains, many clinical applications including elastography involve large strains rendering linear models inaccurate for tissue simulation. Among existing nonlinear models, the Veronda-Westman model has gained much interest because of its exponential form that is consistent with soft tissue mechanical response. However, in elastography where the spatial distribution of this model's parameters must be determined by solving an inverse problem, the exponential form poses serious challenges such as convergence and computation time. To solve the inverse problem, previous methods involved using time-demanding optimization/regularization routines. In this work, we propose a novel technique that does not involve optimization/regularization.

Keywords: Soft tissues, elastography, hyperelasticity, Veronda-Westman, inverse problem

1 INTRODUCTION

According to statistics, cancer is the second leading cause of death worldwide. There are many types of cancer humans can develop among which breast cancer is the second most common type in women [1]. Another common cancer is liver cancer, which is the fifth most common type worldwide, and has the highest mortality rate of 97% in diagnosed patients [2]. The most important factor in the treatment of all types of cancers is early detection and diagnosis. If the cancerous tissue is diagnosed at early stages, there is a greater chance of treating it with little risk to the patient's health. Currently, medical imaging is the most common way to detect cancerous lesions. While sufficient for detecting pathology, many conventional imaging modalities (e.g. CT, MR) suffer from low specificity. This means that although the presence of a tumor within the tissue can be detected, such imaging modalities provide very limited information about the type of the detected abnormality, and most importantly whether it is malignant. To address this, researchers have proposed several methods including elastography, which images tissue stiffness. This is an important development, as data indicate that various pathological tissues exhibit different stiffness characteristics.

2 THEORY

2.1 Elastography

Elastography is an imaging technique in which tissue stiffness is imaged and used to detect or classify tumors. In this work, we focus on the classification capability of elastography, as the presence of tumor can be ascertained using other conventional imaging modalities. Elastography was first introduced by J. Ophir *et al* [3]. The basic idea behind elastography is the fact that tissue pathological changes often trigger substantial stiffness changes. The

core of elastography techniques is their inverse problem of stiffness parameter reconstruction. Reconstruction techniques are based on elasticity constitutive models that are selected to model the forward problem. These are divided into linear and nonlinear (hyperelastic) models. Linear elasticity assumes that the relationship between stress and strain is linear, and uses two parameters (Young's modulus and Poisson's ratio) to describe the mechanical behavior of tissue. However, given that most soft tissues exhibit nonlinear characteristics under the mechanical stimulation of elastography procedures, we employ a hyperelastic formulation. Moreover, tissue hyperelastic parameters can be used for cancer diagnosis. The constitutive relationship of incompressible hyperelastic materials is as follows:

$$S = \frac{2}{J} DEV \left[\left(\frac{\partial U}{\partial \bar{I}_1} + \bar{I}_1 \frac{\partial U}{\partial \bar{I}_2} \right) \bar{B} - \frac{\partial U}{\partial \bar{I}_2} \bar{B}\bar{B} \right] \quad (1)$$

In this equation S is the deviatoric stress, DEV indicates the deviatoric part, and U is a strain energy function. Other parameters (\bar{I}_1 , \bar{I}_2 , \bar{I}_3 , \bar{B} and $\bar{B}\bar{B}$) are functions of displacements and can be calculated using the acquired displacement data.

2.2 Veronda-Westman Model

One of the best nonlinear models in terms of providing a very close fit to typical stress-strain curves of soft tissues is the Veronda-Westman model, which was originally introduced in 1970 [4]. It has been recently used by several researchers in modeling soft tissues [5, 6, 7]. This model has an exponential form, and uses three parameters (C_1 , C_2 and C_3) to describe tissue nonlinear behavior:

$$U = C_1 [e^{C_3(\bar{I}_1-3)} - 1] + C_2(\bar{I}_2 - 3) \quad (2)$$

3 METHODS

3.1 Problem Definition

We need to solve an inverse problem where the tissue displacement data is available to reconstruct the tissue hyperelastic parameters. To form the inverse equations we substitute the Veronda-Westman energy function (Equation (2)) in Equation (1) to obtain:

$$S = \frac{2}{J} DEV \left[(C_1 C_3 e^{(C_3(\bar{I}_1-3))} - C_2 \bar{I}_1) \bar{B} + C_2 \bar{B}\bar{B} \right] \quad (3)$$

This equation is nonlinear in terms of C_1 and C_3 ; therefore, simple matrix-based system inversion is not possible. Previous inversion techniques have used optimization and/or regularization steps for parameter reconstruction [5, 6]. The novelty of this work is the development of a new approach where no optimization or regularization is necessary. With this approach, the complexity and computation time are reduced drastically.

3.2 Novel Inversion Technique

The main idea of the proposed approach is to use an approximation to the exponential term of the Veronda-Westman energy function. Having this approximation, we use a change of variables technique, which changes the system of equations into an equivalent linear system of equations in terms of the new variables. Hence, we solve this linear system for the new variables. Finally, we determine the unknown C_1 , C_2 and C_3 parameters using the obtained new variables. The whole procedure consists of three steps. In each step, one of the unknown parameters of C_2 , C_3 and C_1 is reconstructed.

In the first step, we use the first three terms of Taylor's series to expand the exponential term, and then, to form the described linear system of equations. Since we used a gross approximation of the exponential term, we obtain rough estimations of C_1 and C_3 by solving this equivalent system; however, as Equation (3) is linear in terms of C_2 , we find the accurate value of C_2 in this step.

In the second step, we improve the estimated C values, as we have an estimation of C_3 from the previous step, which enables us to obtain a much better approximation to the exponential term. One way to obtain such an approximation is to use the *polyfit* function in MATLAB, which uses the least squared errors measure to find the closest polynomial approximation to a given function. Fig. 1 shows the higher accuracy of the *polyfit* approximation compared to the Taylor's series expansion. Hence, once again the procedure in the first step is repeated, this time to find the accurate value of C_3 .

To find C_1 , Equation (3) can be solved directly for C_1 based on the obtained C_2 and C_3 . This whole process takes less than a second to finish since no regularization or optimization steps are included.

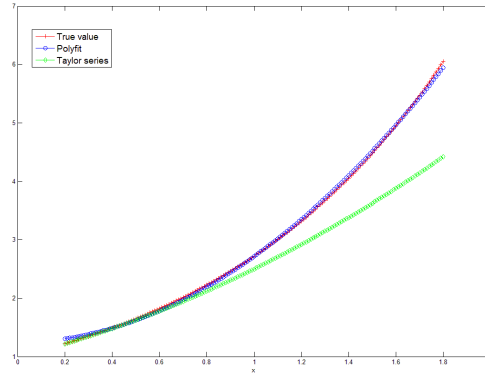


Figure 1. Comparison of two different approximations of polynomial fitting and Taylor's series

4 NUMERICAL PHANTOM VALIDATION

4.1 2-D Numerical Phantom

In this validation, we consider a breast undergoing ultrasound (US) elastography. We created a Finite Element (FE) model of a 2-D breast phantom (Fig. 2) to validate the proposed method. This phantom consists of three different layers: an elliptical inclusion, middle and outer layers which represent the tumor, fibroglandular and adipose tissues, respectively. We use ABAQUS software to perform a FE analysis where we apply 30% of compression to simulate the compression applied by a US imaging probe. To solve the inverse problem we assume that the hyperelastic parameters of the normal (adipose and fibroglandular) tissues are available [8]. Thus, we follow a constrained reconstruction procedure similar to the one presented by Mehrabian *et al* [6].

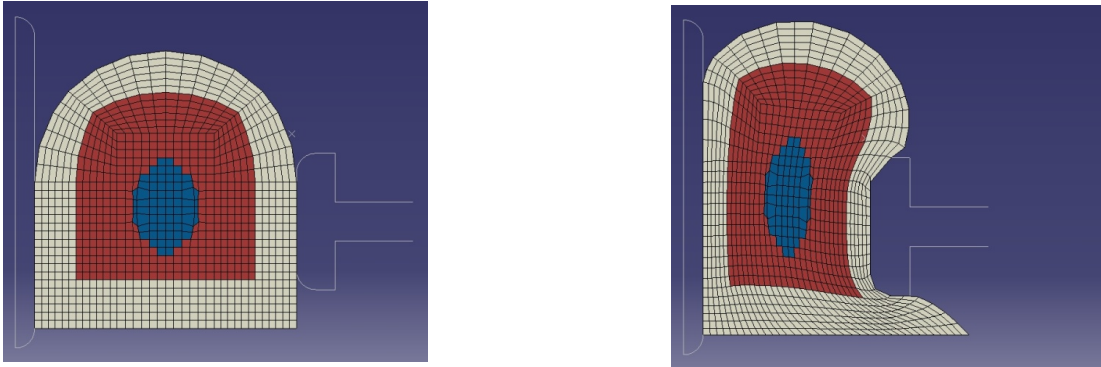


Figure 2. FE model of the phantom before applying deformation, and after applying ultrasound probe compression

4.2 3-D Numerical Phantom

In this example, a 3-D breast phantom undergoing MR elastography is studied. This phantom is similar to the 2-D counterpart. It has the same three layers (adipose and fibroglandular tissues, and a tumor). Different views of this phantom are shown in Fig. 3.

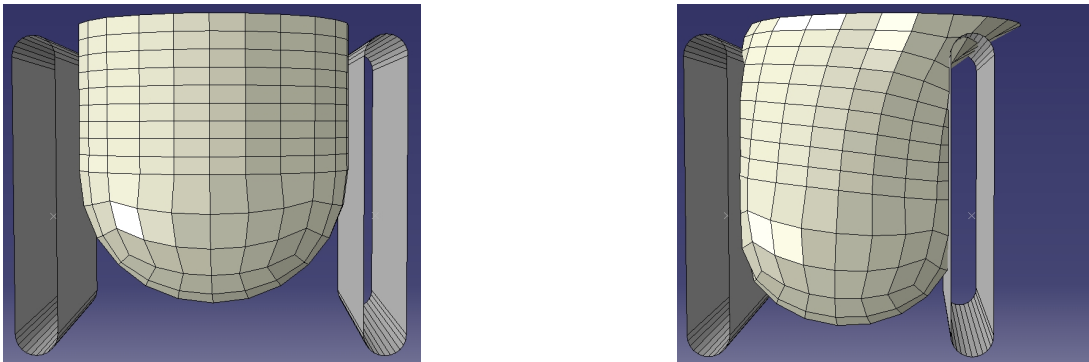


Figure 3. FE model for 3-D breast mimicking phantom in pre-compression and post-compression states

Again, we solve for the hyperelastic parameters of the tumor tissue assuming known parameters for the normal tissues. We use the same reconstruction algorithm here. The only difference is that the equations are formed considering three dimensions instead of a 2-D framework.

5 RESULTS

We tested the method on both 2-D and 3-D phantoms for a wide range of hyperelastic parameters on 100 different sets. These 100 sets of parameters were randomly selected such that they covered the entire range of hyperelastic parameters that O'Hagan [9] has reported for Veronda-Westman model of real breast tumor samples. As an indication, we have included reconstruction results for five sets of those 100 samples here. These five sets are shown in Table 1, and corresponding reconstruction results for 2-D and 3-D cases are shown in Tables 2 and 3, respectively. For the 2-D phantom, the maximum error we obtained was less than 4%, and the average error was 1.93%. For the 3-D phantom, we observed a maximum error of less than 2%, and an average error of 1.11%.

Table 1. Parameters used to construct both 2-D and 3-D phantoms

Test No.	Real Parameters (C_1 and C_2 are in kPa, C_3 is unitless)								
	Adipose tissue			Fibroglandular tissue			Tumor tissue		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
1	2.79	-1.12	1.11	5.38	-2.29	2.58	8.20	-2.15	3.39
2	3.28	-1.51	2.23	5.26	-3.46	3.13	9.37	-2.33	4.25
3	4.59	-1.60	1.29	7.67	-2.22	1.94	10.13	-3.66	2.81
4	3.81	-0.49	1.54	5.83	-3.25	2.31	11.05	-2.41	3.72
5	5.68	-1.47	1.34	8.25	-4.29	2.30	13.68	-3.72	3.38

Table 2. Reconstructed tumor parameters for the 2-D phantom

Test No.	Reconstructed Parameters (C_1 and C_2 are in kPa, C_3 is unitless)								
	C_1 (true)	C_1 (solved)	Error (%)	C_2 (true)	C_2 (solved)	Error (%)	C_3 (true)	C_3 (solved)	Error (%)
1	8.20	8.29	1.10	-2.15	-2.09	2.79	3.39	3.35	1.18
2	9.37	9.54	1.81	-2.33	-2.27	2.58	4.25	4.09	3.76
3	10.13	10.24	1.09	-3.66	-3.63	0.82	2.91	3.01	3.43
4	11.05	11.22	1.54	-2.41	-2.36	2.07	3.72	3.60	3.23
5	13.68	13.37	2.27	-3.72	-3.58	3.76	3.38	3.42	1.18

Table 3. Reconstructed tumor parameters for the 3-D phantom

Test No.	Reconstructed Parameters (C_1 and C_2 are in kPa, C_3 is unitless)								
	C_1 (true)	C_1 (solved)	Error (%)	C_2 (true)	C_2 (solved)	Error (%)	C_3 (true)	C_3 (solved)	Error (%)
1	8.20	8.06	1.71	-2.15	-2.18	1.40	3.39	3.44	1.47
2	9.37	9.27	1.07	-2.33	-2.32	0.43	4.25	4.29	0.94
3	10.13	10.01	1.85	-3.66	-3.62	1.09	2.91	2.86	1.72
4	11.05	11.03	0.18	-2.41	-2.44	1.24	3.72	3.72	0.00
5	13.68	13.54	1.02	-3.72	-3.67	1.34	3.38	3.40	0.59

6 DISCUSSION AND CONCLUSION

Elastography is a powerful technique that can be used to determine the types of breast tumors. While linear elastography lacks the capability to model soft tissues because of their nonlinear behavior, hyperelastic models are better suited for this purpose. Veonda-Westman is a hyperelastic model that has been used recently to model soft tissues, especially breast tissue. Its exponential form makes it capable of modeling nonlinear behavior of soft tissues highly accurately; however, its corresponding hyperelastic parameter reconstruction problem leads to a nonlinear system of equations.

Previously proposed techniques involve regularization and/or optimization, which are very time-consuming and are not guaranteed to converge. In this work, we presented a novel solution to this nonlinear system which is much faster and yet reasonably accurate. This solution is straight-forward, as it does not use regularization or optimization. The technique was validated using 2-D and 3-D phantom studies where its performance in reconstructing tissue hyperelastic parameters was demonstrated. These studies indicated reconstruction errors of less than 4% and 2% in the 2-D and 3-D phantoms, respectively. In terms of speed, the proposed method reduced the computation time by 1/9 compared to Gokhale's work [7], and by 1/3 compared to Mehrabian's work [8].

Future work will involve experimental validation of the proposed technique using a tissue mimicking phantom. The phantom will be constructed using PVA-C which has been used widely in modeling soft tissues because of its nonlinear behavior. The phantom's displacement data will be acquired using an ultrasound system.

References

- [1] World Health Organization International Agency for Research on Cancer, "World Cancer Report 2008".
- [2] World Health Organization International Agency for Research on Cancer, "World Cancer Report 2003".
- [3] J. Ophir, I. Cespedes, H. Ponnekanti, Y. Yazdi, and X. Li, "Elastography: A quantitative method for imaging the elasticity of biological tissues," *Ultrasonic Imaging*, vol. 13, pp. 111-134, 1991.
- [4] D. R. Veronda and R. A. Westmann, "Mechanical characterization of skin-Finite deformations," *Journal of Biomechanics*, vol. 3, pp. 111-122, IN9, 123-124, 1970.
- [5] Gokhale N. H., Barbone P. E., Oberai A. A., "Solution of the nonlinear elasticity imaging inverse problem: The compressible case," *Inverse Problems*, vol. 24, (2008).
- [6] H. Mehrabian and A. Samani, "Constrained hyperelastic parameters reconstruction of PVA (Polyvinyl Alcohol) phantom undergoing large deformation," *SPIE*, Vol. 7261, 72612G.
- [7] H. M. Yin, L. Z. Sun, G. Wang, and M. W. Vannier, "Modeling of Elastic Modulus Evolution of Cirrhotic Human Liver," *IEEE Transactions on biomedical engineering*, no. 10, vol. 51, 2004.
- [8] Samani A., Plewes D. B., "A method to measure the hyperelastic parameters of *ex vivo* breast tissue samples," *Physics in Medicine and Biology*, vol. 49, pp. 4395-4405 (2004).
- [9] J. J. O'Hagan, A. Samani, "Measurement of the hyperelastic properties of 44 pathological *ex vivo* breast tissue samples," *Physics in Medicine and Biology*, vol. 54, pp. 2557-2569.

Evaluating a motor unit potential train using cluster validation methods

Hossein Parsaei¹ and Daniel W. Stashuk

Systems Design Engineering, University of Waterloo, Waterloo, Canada;

ABSTRACT

Assessing the validity of motor unit potential trains (MUPTs) obtained by decomposing a needle-detected electromyographic (EMG) signal is a crucial step in using these trains for quantitative EMG analysis. In general, for MUPT validation a train is assessed using the shapes of its motor unit potentials (MUPs) and the motor unit firing pattern it represents. Here, two methods to assess the validity of a given MUPT using its MUP shape information are presented. These methods are based on the gap statistic and jump algorithms presented for estimating the number of clusters in a dataset. They evaluate the shapes of the MUPs of a MUPT to determine whether it represents the activity of a single MU (i.e. it is a valid MUPT) or not. Evaluation results using both simulated and real data show the gap statistic method is more accurate than the jump method in correctly categorizing a train. The accuracy of the gap statistic method was 92.3% for simulated data and 93.8% for real data while accuracy for the jump method was 88.3% and 91.0%, respectively. The results are encouraging and suggest that using these methods can improve EMG signal decomposition results, and facilitate automatic MUPT validation.

Keywords: EMG signal decomposition, motor unit potential train, motor unit potential wave shape, motor unit potential train validation, cluster validation.

1 INTRODUCTION

An electromyographic (EMG) signal is simply the superposition of the electrical activity detected by an electrode of the motor units (MUs) that are active during muscle contraction. A motor unit is a single α -motor neuron, its axon and all the connected muscle fibers. MUs are repetitively active during sustained voluntary contraction and generate trains of motor unit potentials (MUPs), each of which is called a motor unit potential train (MUPT). Therefore, an EMG signal is the summation of MUPTs and when detected using suitable electrodes reflect the characteristics of the muscle from which it was detected.

Recent development in computer technology, signal processing and pattern recognition techniques have provided researchers and engineers with opportunities to develop new techniques for extracting valuable information regarding a contracting muscle from an EMG signal detected from this muscle. One of these techniques is EMG signal decomposition which is the process of resolving a composite EMG signal into its constituent MUPTs. This is implemented by employing digital signal processing and pattern recognition techniques in four steps: signal preprocessing, signal segmentation and MUP detection, feature extraction, clustering of detected MUPs, and supervised classification of detected MUPs [1]. The first step is to remove background noise and low-frequency information from the detected EMG signal, to shorten the duration of MUPs and decrease MUP temporal overlap, and to sharpen the MUPs and increase discrimination between them. The second step is to section the signal into segments containing possible MUPs that were generated by active motor units and contribute significantly to the detected EMG signal. The detected MUPs are represented by a feature vector in the third steps and finally are sorted into MUPTs using clustering and/or supervised classification techniques. The obtained MUPTs provide information regarding the temporal behavior and morphological layout of the generating MUs. This information can assist with the diagnosis of various neuromuscular diseases and the study of motor unit control, and lead to a better understanding of healthy, pathological, ageing or fatiguing neuromuscular systems [1-5]. However, this is achieved only when this information is valid. In fact, before using decomposition results and the MUP shape and MU firing pattern information for either clinical or research purposes the validity of the extracted MUPTs needs to be confirmed. Although many EMG signal decomposition methods have been developed, automatic validation of the extracted MUPTs has not been investigated in detail.

¹ Corresponding author. E-mail: hparsaei@engmail.uwaterloo.ca, Telephone: +1(519) 888- 4567 Ext.33746.

To date, a decomposition-created MUPT is evaluated qualitatively by an expert operator using the shapes and occurrence times of the MUPs assigned to the train [1,5,6]. MUP shape-based validation of a MUPT is conducted by assessing the raster/shimmer plots of the MUPs assigned into this train to determine if their wave shapes are consistent or not. In fact, if the shapes of the MUPs assigned to a given MUPT are consistent, one can conclude that this MUPT represents the MUPs of a single MU and is valid; otherwise it is an invalid train. Firing pattern-based validation of a MUPT is made by evaluating its inter-discharge interval (IDI) histogram [1] and the instantaneous firing rate of the corresponding MU versus time. The MU discharges corresponding to a valid MUPT occur at regular intervals, while an invalid MUPT that does not represent the activity of a single MU and contains many false positive errors will have large variations in its firing rate plot and will not have a Gaussian like IDI distribution. A train is considered valid if it satisfies both temporal and shape criteria.

Although qualitative evaluation of MUPTs does not depend on the decomposition algorithm used or the signal decomposed, it does depend on operator experience and skill. Moreover, it is time consuming and hence cannot be practically completed in a busy clinical environment. To overcome these issues, methods need to be developed to automatically assess MUPT validity. Parsaei et al. [7] developed a supervised method for validating a MUPT using its firing pattern information. In these methods, ten features of the MUPT firing pattern are extracted and then fed to two supervised classifiers and to a linear model to determine whether they represent the firings of a single MU or the merged activity of more than one MU, and if it is a single train whether the estimated levels of false positive and false negative errors in it are acceptable or not. Here, automatic MUP-shape based validation of a MUPT is explored. Two methods based on cluster validation concepts are proposed to automatically validate MUPTs using their MUP shape information. They evaluate the shapes of the MUPs of a MUPT to determine whether they are consistent or not. If a train passes this test, it can be concluded that it represents the activity of a single motor unit and hence is valid. The composition of these methods, their objectives and how they were evaluated using both simulated and real data are presented below.

2 VALIDATING A MUPT USING MUP SHAPE INFORMATION

To convey the concept of assessing validity of a MUPT using its MUP shape information, two examples are provided in Figure 1. The left column shows a valid MUPT and the right column shows an invalid train. The valid train was obtained from decomposing a simulated EMG signal. The invalid train was created by merging two valid MUPTs. As shown, the shapes of MUPs assigned to the valid MUPT are consistent while that of the invalid train are inconsistent. The shapes of MUPs in the invalid train are different for samples 11 to 25. The goal of developing a MUPT validation method is to perform this assessment automatically during or after decomposition. The advantages of using such an algorithm during EMG signal decomposition is that detecting invalid trains during decomposition can improve decomposition accuracy by improving estimation of the number of MUPTs, their MU firing pattern statistics and MUP templates.

On the whole, the process of EMG signal decomposition can be considered as a clustering problem because neither the number of MUPTs (i.e. clusters) nor the labels of the MUPs are known in advance. During EMG signal decomposition, detected MUPs are clustered into groups called MUPTs. Therefore, shape-based validation of a MUPT can be considered a cluster validation problem and the decision to be made is whether a decomposition-created MUPT represents one cluster in terms of the shapes of the assigned MUPs or not. If the MUPs of a given MUPT are homogeneous in terms of their shapes, they will represent one cluster and hence it can be concluded

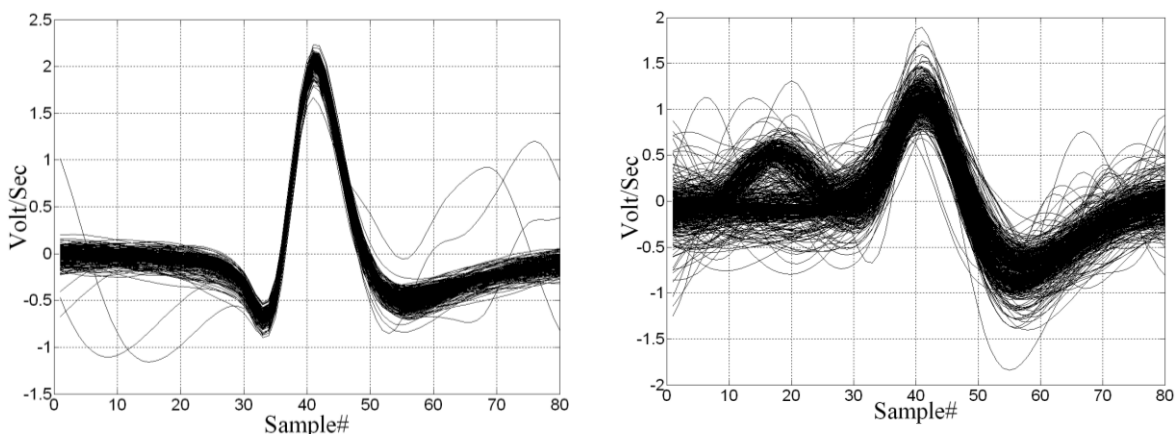


Figure 1. A valid MUPT (left) and a simulated invalid MUPT (right).

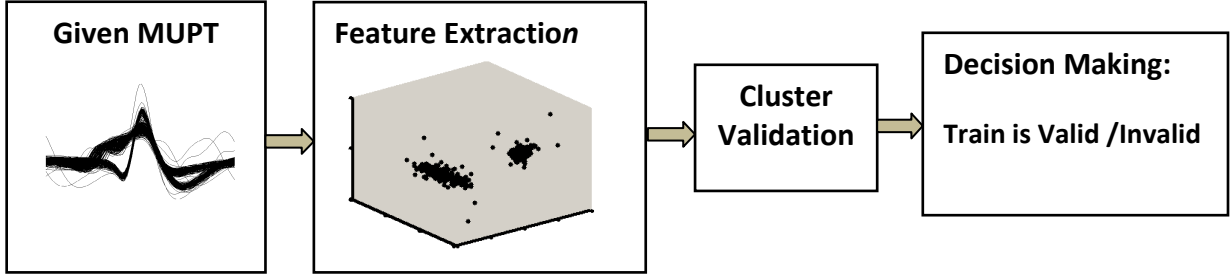


Figure 2. Overview of an automatic MUPT validation algorithm using its MUP shape information.

that this MUPT represents the MUPs of a single motor unit and is valid; otherwise this is an invalid train. Figure 2 summarizes the principal steps of a MUP shape-based MUPT validation algorithm.

In this work, two algorithms based on the gap statistic [8] and the jump methods [9] have been developed for this purpose. Both algorithms include two steps: feature extraction and cluster validation. For features extraction, two techniques were employed. For the first feature space, 80 time samples of the MUPs filtered by a low-pass differentiator (LPD) filter and then centered on the position of their peaks were used. The LPD filtered samples were used instead of unfiltered samples because they discriminate between the MUPs generated by different motor units better than the raw samples. For the second feature space, a reduced number of uncorrelated features selected via principle components analysis (PCA) were used. The MUPs of a given train were represented by the first T principal components that represent $\beta\%$ of the data variance. Given the features extracted, the validity of the given train was assessed using either the gap statistic or jump method.

Both the gap statistic and jump methods were mainly developed for estimating the optimum number of groups (\hat{K}) in a data set, but they can also be used for evaluating a single cluster such that they are classified as global cluster validation methods. In general, global cluster validation methods compare two/more clustering results for deciding which one fits the data best. For finding \hat{K} in a given data set via these methods, the quality of clustering results is measured by a criterion and then is optimized as a function of the number of clusters when the entire data set is clustered into K groups for $K=1, 2, 3, \dots, K^*$. Where K^* is the maximum possible number of clusters in the given data set. It is clear that \hat{K} will be the value of K for which this criterion is optimal. In general, the criteria are defined based on the assumption that a cluster is compact and well separated from other clusters. Hence, they are often defined based on the within-cluster scatter dispersion (W_K) and between-cluster scatter variability (B_K). W_K is given by [8]:

$$W_K = \sum_{i=1}^K \sum_{\underline{x} \in c_i} (\underline{x} - \underline{m}_i)(\underline{x} - \underline{m}_i)^T \quad (1)$$

where \underline{m}_i is the sample mean for the N_i members of the i th cluster.

The gap statistic [8] method estimates the number of clusters by comparing $\log(W_K)$ to its expected value (i.e., $E\{\log(W_K)\}$) estimated using an appropriate null reference distribution of the given data set. Defining the gap statistic as $\text{Gap}_K = E\{\log(W_K)\} - \log(W_K)$, the best value for \hat{K} corresponds to the first local maximum of Gap_K . Tibshirani et al. [8] proposed two methods to generate a reference data set for the gap statistic method. In the first method, it is sampled uniformly from the range of the observed values for each feature. In the second method, the reference data set is also sampled uniformly but here over a box aligned with the principal components of the data. The gap statistic method based on the first method is known as the gap/uni method and that based on the second method is known as the gap/pc method. The former is simpler than the latter but may not be as accurate because the gap/pc considers the shape of the cluster in generating the reference data set and hence has a better estimation of $E\{\log(W_K)\}$.

The jump method [9] applies an appropriate transformation to the curve of W_K and then determines the largest jump in the transformed curve. The value of K associated with this jump is considered as the best value for \hat{K} . Sugar and James [9] proposed the following transformation for W_K

$$W_K^* = W_K^{-Y} \quad (2)$$

where Y is the transformation power. A typical value for Y is $d/2$, where d is the dimension of the feature space. Defining $W_0^* = 0$, the jump index (J_K) is given by $J_K = W_K^* - W_{K-1}^* = W_K^{-Y} - W_{K-1}^{-Y}$; $K=1, 2, \dots, K^*$. Given J_K , $\hat{K} = \text{argmax}_K(J_K)$. The theoretical results provided by Sugar and James [9] show that $d/2$ is the best value for Y only when the data has a multivariate independent Gaussian distribution. For the cases that this assumption is not valid, they suggest trying several values of Y to find the best value for this parameter.

Table 1. The parameters and accuracy of the four methods studied for validating a MUPT using its MUP shape information applied to simulated and real data. **VasV** stands for the valid MUPTs classified as valid, **Iv as Iv** represents the invalid MUPTs recognized as invalid, and **Acc** represents the total accuracy.

Method	Parameters	Simulated Data			Real Data		
		V as V %	Iv as Iv %	Acc %	Vas V %	Iv as Iv %	Acc %
Gap statistic	-	83.0±0.4	98.2±0.3*	90.6±0.3	89.2	98.3*	93.8
PGS	$\beta = 90$	92.0±0.6*	93.3±0.3	92.7±0.3*	93.1	97.9	95.5*
Jump	$Y = 3$	81.5±0.8	90.5±0.6	86.0±0.5	87.9	94.0	91.0
PJ	$\beta = 50; Y = 2$	86.2±0.6	90.4±0.6	88.3±0.4	89.1	93.7	91.4

In order to assess the validity of a given MUPT using the jump method, the train is split into $K=1$ to K^* sub trains using a K-means algorithm and then if $\max(J_K) = J_1$, the given train is labeled a valid train otherwise it is labeled an invalid train. If the gap statistic method is to be used for validating this train, the same procedure will be repeated but the gap criterion will be used for cluster validation proposes. However, it is shown that for a given data set including compact and well separated clusters W_K decreases monotonically as the value of K increases, but when K reaches the true number of clusters (i.e., \hat{K}) this decrease becomes smoother. Thus, $\text{Gap}_K < \text{Gap}_{\hat{K}}$ when $K < \hat{K}$ and $\text{Gap}_K > \text{Gap}_{\hat{K}}$ when $K > \hat{K}$. In other words, if $\text{Gap}_1 > \text{Gap}_2$, $\max(\text{Gap}_K)$ will be equivalent to the gap value for $K=1$. Consequently, since the goal is to determine whether a MUPT represents one cluster in terms of the assigned MUP shapes or not, only Gap_K for $K=1$ and 2 are sufficient. Therefore, for the gap statistic-based MUPT validation, the algorithm is only run for $K=1$ and 2 and if $\text{Gap}_1 > \text{Gap}_2$ the MUPT under question is flagged a valid train otherwise it is classified an invalid train. This decreases the algorithms processing time and hence makes it practical for clinical applications. For the jump-based MUPT validation, however, the algorithm must be run for $K=1,2,3,\dots, K^*$ because even if $J_1 > J_2$ the is no guarantee that $\max(J_K) = J_1$ (i.e., $\hat{K}=1$).

3 RESULTS AND DISCUSSION

The effectiveness of the developed method was studied using both simulated and real data. In total four methods, as listed in Table 1 were evaluated. In this Table, PGS and PJ stand for the gap statistic and jump methods when the features used were selected using PCA, respectively.

For simulated data, 261 EMG signals each of 30s length with different levels of intensity (24-93 pps), MUP shape stability (with jitter values from 50-150 μ s) and IDI variability (CV from 0.10-0.45) were generated using an EMG signal simulator [10]. These data allowed the performance relative to signal intensity, number of trains and MUP shape variability to be studied. The simulated signals were decomposed (using the DQEMG software [11]) and the resulting MUPTs visually assessed to determine valid and invalid MUPTs. Additional invalid trains were generated by merging valid MUPTs (up to 4) randomly selected from each signal. Additional valid trains were generated by selecting valid MUPTs with greater than 100 MUPs and randomly splitting them into sub trains of at least 50 MUPs. In total 36000 MUPTs (18000 valid and 18000 invalid trains) were tested. The number of merged trains generated was 234778, but only 18000 were randomly selected to have equal class sizes. Out of the 18000 selected merged trains, 60% include MUPs of two valid MUPTs, 30% include MUPs of three valid MUPTs, and 10% include MUPs of four valid MUPTs. This data set was divided into 30 subsets each containing 600 valid and 600 invalid trains.

For real data, EMG signals provided by M. Nikolic of Rigshospitalet, Copenhagen, Denmark [12] were used. These signals were detected from normal, myopathic and neuropathic muscles using a standard concentric needle electrode during constant low level voluntary contractions. The same analysis as with the simulated data was done on these signals. This dataset includes 3130 MUPTs (1565 valid and 1565 invalid trains).

Before evaluating the four considered methods using the provided simulated and real data, the best values for their user defined parameters were determined empirically using one of the thirty subsets of the simulated data described above. In using the gap statistic method, the gap/pc was used because it outperformed the gap/uni method. For the other methods, the values used for their parameters are listed in the second column of Table 1.

The average accuracy of the developed method in determining valid MUPTs and invalid MUPTs for both the simulated and real data sets are summarized in Table 1. The accuracy here is defined as percentage of the number of correctly classified MUPTs. For example, in the column presenting the results for valid trains, the accuracy represents the percentage of the number of examined valid train labeled as valid by the studied algorithm. The numbers given for simulated data are obtained by testing each method using the thirty different data sets described above. For each column, in the simulated data category the best method as determined by a t-test using a 5% significance level are indicated by '*'. Based on these results, the PGS method is the best algorithm because it is

the most accurate in terms of overall accuracy and labeling valid trains correctly. In classifying invalid MUPTs the gap statistic is the best performer, but the probability of error of this method for valid trains is high. It is 0.17 on average which causes duplication of MUPTs during EMG signal decomposition. The PGS method is the second most accurate with an accuracy of 93% which is acceptable in a practical sense. The results obtained using the real data support the conclusion drawn from the simulated data results that the PGS method is the best algorithm. All four methods studied performed better using the real dataset than the simulated dataset because the variability of the MUPs in the real dataset are lower than those of the simulated EMG signals and also the MUPs created by different MUs in the real dataset are less similar than those in the simulated signals. Most of the valid trains recognized as invalid are trains with highly variable MUP shapes caused by either high numbers of superimpositions (for the signals with high intensity) or very high jitter (around 150 μ s). Therefore, the accuracy of this method in determining valid MUPTs will be higher for trains provided by EMG decomposition algorithms that resolve superimposed MUPs. Most of the invalid trains that labeled incorrectly are trains with very similar MUP shapes. Such trains are hard to assess using only shape information, but firing pattern information can assist with label them correctly [7]. Nevertheless, the obtained results are encouraging and suggest that using these methods can facilitate automatic validation of a MUPT extracted from a decomposed EMG signal. It can also improve EMG signal decomposition results, by obtaining more accurate estimates of the number of MUPTs, and the MUP template and MU firing pattern statistics of each MUPT.

Acknowledgements

The authors gratefully acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- [1] D.W. Stashuk, "EMG signal decomposition: how can it be accomplished and used? " *J. of Electromyography and Kinesiology*, **1**, pp. 151-173, 2001.
- [2] A. Fuglsangfrederiksen, "The role of different EMG methods in evaluating myopathy," *Clinical Neurophysiology*, **117**, pp. 1173-1189, 2006.
- [3] E. Sålberg and B. Falck, "The role of electromyography in neurology," *Electroencephalography and Clinical Neurophysiology*, **103**, pp. 579-598, 1997.
- [4] C. J. de Luca, R.S. LeFever, M.P. McCue, and A.P. Xenakis, "Control scheme governing concurrently active human motor units during voluntary contractions," *J. of Physiology*, 329, pp. 129-142, 1982.
- [5] K.M. Calder, D.W. Stashuk, and L. McLean, "Physiological characteristics of motor units in the brachioradialis muscle across fatiguing low-level isometric contractions," *J. of Electromyography and Kinesiology*, **18**, pp. 2-15, 2008.
- [6] S.G. Boe, D.W. Stashuk, W.F. Brown, and T.J. Doherty, "Decomposition-based quantitative electromyography: Effect of force on motor unit potentials and motor unit number estimates," *Muscle & Nerve*, **31**, pp. 365-373, 2005.
- [7] H. Parsaei, F. Jahanmiri Nezhad, D.W. Stashuk, and A. Hamilton-Wright, "Validation of motor unit potential trains using motor unit firing pattern information," in 31st Annu. Int. Conf. of IEEE EMBS Minnesota, USA, pp. 974-977, 2009.
- [8] R. Tibshirani R, G. Walther G, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. of the Royal Statistical Society*, **63**, pp. 411-423, 2000.
- [9] A.C. Sugar, and G. M. James, "Finding the number of clusters in a data set: an information theoretic approach," *J. of the American Statistical Association*, **98**, pp.750-763, 2003.
- [10] A. Hamilton-Wright and D.W. Stashuk, "Physiologically based simulation of clinical EMG signals," *IEEE Trans. on Biomed. Eng.*, **52**, pp.171-183, 2005.
- [11] D. W. Stashuk, "Decomposition and quantitative analysis of clinical electromyographic signals," *Med. Eng. & Phys.*, **21**, pp.389-404,1999
- [12] M. Nikolic, "Detailed Analysis of Clinical Electromyography Signals EMG Decomposition, Findings and Firing Pattern Analysis in Controls and Patients with Myopathy and Amyotrophic Lateral Sclerosis," Ph.D. dissertation, Faculty of Health Science, University of Copenhagen, 2001.

***In silico* study of the interaction of the Myelin Basic Protein C-terminal α -helical peptide with DMPC and mixed DMPC/DMPE lipid bilayers**

Kyrylo Bessonov ¹

University of Guelph, 50 Stone Road East, Guelph, Canada

ABSTRACT

Biological membranes continue to be extensively investigated in different ways. This paper presents the benefits of Molecular Dynamics (MD) approaches to study the properties of biological membranes and proteins using the freely available GROMACS package, in the context of the Myelin Basic Protein (MBP) C-terminal α -helical peptide. A mixed membrane consisting of 2-Dimyristoyl-*sn*-Glycero-3-phosphocholine/1,2-Dimyristoyl-*sn*-Glycero-3-phosphoethanolamine (DMPC/DMPE), and pure DMPC membranes, composed of 188 and 248 lipids, respectively, were simulated for 200 ns at 309 K. The DMPC membrane was approximately three times more fluid compared to the DMPC/DMPE system, with the diffusion coefficients (*D*) being $0.0207 \times 10^{-5} \text{ cm}^2/\text{s}$ and $0.0068 \times 10^{-5} \text{ cm}^2/\text{s}$, respectively. In addition, the 14-residue peptide representing the C-terminal α -helical region of murine Myelin Basic Protein (MBP), with amino acid sequence $\text{NH}_2\text{-A}_{141}\text{YDAQGTLISKIFKL}_{154}\text{-COOH}$, was simulated in both membrane systems for 200 ns. The peptide penetrated further into the DMPC bilayer compared to the mixed DMPC/DMPE bilayer, potentially because of the reduced accessibility of the charged peptide amino acid side chains to the formal positive charge of the amine N atom surrounded by methyl and methylene groups in DMPC, that might have resulted in greater overall peptide mobility [3]. These findings are significant in their implication that membrane composition affects the behavior of MBP, providing further insights into myelin structure. Our preliminary results suggest that local changes in membrane composition (e.g. enrichment in DMPE molecules), as well as, electrostatic nature of primary amino acid sequence could cause localized denaturation / instability of external MBP α -helices possibly augmenting the degradation of myelin in multiple sclerosis (MS), resulting in a subsequent decrease of nerve impulse propagation efficiency.

Keywords: myelin, MBP, GROMACS, Multiple Sclerosis, DMPC, DMPE

1 INTRODUCTION

Myelin Basic Protein (MBP) is an important protein in the central nervous system. The protein is found in various isoforms with a predominant splice isoform of 18.5 kDa in an adult brain. The main physiological role of MBP is to maintain the myelin sheath that wraps around neurons by holding together both cytoplasmic sides of oligodendrocyte membranes, thus facilitating the compaction of the myelin sheath and allowing efficient signal propagation [7].

Recent studies have demonstrated that the severity of MS is correlated with post-translational modifications of MBP, such as citrullination [8]. Due to its central role, MBP is thought to be connected with myelin degradation. MS attacks the myelin-wrapped nerves of the central nervous system. Molecular Dynamics (MD) provides a nice, quick way to study the behavior and interaction patterns of MBP with lipid membranes that could provide insights into molecular details of myelin structure, and pathogenic mechanisms in MS.

The main focus of this article is to provide both a practical and methodological approach to MD using GROMACS [10], as well as to introduce possible applications of such simulations to real biological problems. The supplementary website provides additional information, key files and additional programs that facilitate the setup of MD simulations. Here, the simulations of DMPC and mixed DMPC/DMPE (1:1 ratio) membranes and MBP C-terminal peptide were performed on SHARCNET cluster revealing importance of membrane composition on MBP behavior useful to further knowledge on MS pathogenesis.

¹ Corresponding author. E-mail: kbessonov@uoguelph.ca Website: <http://www.uoguelph.ca/~kbessonov/main.html>

2 METHODS

2.1 Purpose

To investigate how both DMPC and DMPC/DMPE membranes affect the behavior of a 14-residue peptide representing one of three α -helical regions of the classical murine Myelin Basic Protein (MBP). The sequence $^{\text{NH}}_2\text{-A}_{141}\text{YDAQGTLISKIFKL}_{154}\text{-COOH}$ was modeled as 14-residue α -helical peptide using Molecular Operations Environment 2008 (MOE, Chemical Computing Group, Montreal).

2.2 Preparation of DMPC and mixed DMPC/DMPE membranes for MD simulation

The peptide was carefully positioned slightly above the lipid bilayer using both Visual Molecular Dynamics (VMD) and self-written *gro_mover* programs. DMPC and DMPC/DMPE membranes were neutralized with Na^+ and Cl^- ions inserted into the aqueous layer using *genion* and giving an overall system charge of zero, and an overall system pH of 7.0. To prevent overlap between atoms and increase stability of the system, energy minimization (EM) using the steepest descents method was done for both membranes. EM finds the system local potential energy minimum by using a specified force field. EM is usually required to be done before any MD run, because the solvation of lipid membranes in water usually introduces some bad contacts/atom clashes that need to be relaxed before being given kinetic energy (i.e., MD). The assembled membrane systems were simulated on the SHARCNETTM computer cluster using 96 processors for a total trajectory time of 200ns.

2.3 The $^{\text{NH}}_2\text{-A}_{141}\text{YDAQGTLISKIFKL}_{154}\text{-COOH}$ properties

The peptide had an overall +1 charge and displayed 38% hydrophilicity based on its primary sequence (Fig. 1). The negatively-charged aspartic acid (D) residue confers -1 charge to peptide's N-terminus, while two lysine (K) residues ensure a +2 charge at the C-terminal end. The overall peptide pI is 9.6

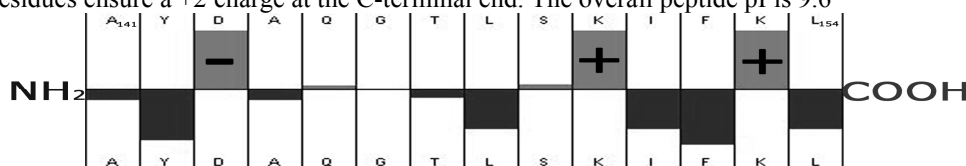


Figure 1: Hydrophilicity plot of the $^{\text{NH}}_2\text{-A}_{141}\text{YDAQGTLISKIFKL}_{154}\text{-COOH}$ showing N- and C-termini. The Y-axis represents the Hopp-Woods hydrophilicity scale while + and - signs refer to amino acid R-side chain charge [11].

3 RESULTS and DISCUSSION

3.1 Measuring DMPC and DMPC/DMPE lipid bilayer parameters

Membrane dynamics simulations provide a powerful means for studying how temperature, protein, cholesterol content, and numerous other parameters affect membrane characteristics such as fluidity and lipid velocity.

The use of computer clusters through SHARCNETTM significantly diminished the total trajectory computation time. The assembled DMPC and DMPC/DMPE systems were simulated for 200 ns. The obtained DMPC and DMPC/DMPE trajectory files were analyzed against diffusion coefficients, total kinetic energies, pressure, temperature, and other parameters (see **Table 1**) using *g_energy* [10] and an *InflateGro* Perl script [4]. Other parameters, such as solvent accessibility, were not successfully measured due to *g_sas*'s difficulty [10] in recognition of the hydrophobic parts of the DMPC molecule.

Table 1. Summary of measured parameters for DMPC and DMPC/DMPE membrane systems used in MD simulations.

Membrane characteristic	DMPC	DMPC/DMPE
Total number of atoms	62,613	30,416
Total number of lipid molecules	248	94/94 (188)
Diffusion coefficient (D) 10 ⁻⁵ cm ² /s	0.0207	0.0068
Kinetic Energy (J/mol)	161,870	78,602
Total Energy (J/mol)	-891,474	-477,439
Heat Capacity Cv (J/mol*K)	12.4721	12.4724
Temperature (K)	309	309
Pressure (bar)	1.66	1.097
Average Area per lipid (Å ² /lipid)	67.26	55.24
Membrane Thickness (Å)	33.9 – 35.72	35.8-38

Diffusion coefficients (D) describe the mobility of the molecules. The higher D values are indicative of greater mobility. The simulation results (**Table 1**) indicate that a DMPC bilayer has three times greater fluidity as compared to the mixed DMPC/DMPE bilayer at 309 K, as suggested by the diffusion coefficients (D) of $0.0207 \times 10^{-5} \text{ cm}^2/\text{s}$ and $0.0068 \times 10^{-5} \text{ cm}^2/\text{s}$, respectively. The difference in fluidity between both membranes could be partially explained by the difference in the density of lipid packing. The DMPC membrane was found to be more loosely packed as compared to the DMPC/DMPE membrane system, with average areas per lipid of 67 \AA^2 and 55 \AA^2 , respectively. It is expected that membranes with a higher density of lipid packing will restrict movement of freely diffusible molecules such as peptides.

Membrane thickness was calculated by labeling phosphate atoms of lipid molecules on opposite sides of the bilayer with the help of VMD 1.8.6 software. The thickness did not change significantly during simulation, indicating stability of the membrane. Accurate membrane thickness determination was hindered by constant random lipid motion in both bilayers (data not shown).

3.2 Simulation of 14-residue MBP C-terminal peptide [$^{\text{NH}}_2\text{-A}_{141}\text{YDAQGTLISKIFKL}_{154}\text{-COOH}$]

In addition to bare membrane simulations, the simulation of the 14-residue long MBP C-terminal peptide in DMPC and mixed DMPC/DMPE bilayers was done for the first time. Interesting trends of the two systems related to membrane composition and protein secondary structure preservation were observed over the course of the 200 ns simulation. The depth of penetration and α -helical structure stability were successfully measured (data not presented here). Overall, the peptide penetrated further into the DMPC membrane as compared to the mixed DMPC/DMPE membrane. Helical secondary structure retention was stronger in the DMPC bilayer system (**Fig. 2**). This might be due to DMPC and DMPE N atom formal positive charge distribution and accessibility differences as explained in Section 3.3. Thus, the electrostatic environment of the two membranes might partially dictate stability of the peptide amongst other factors (i.e., localized pH, lipid-peptide thermodynamics, lipid density) [12].

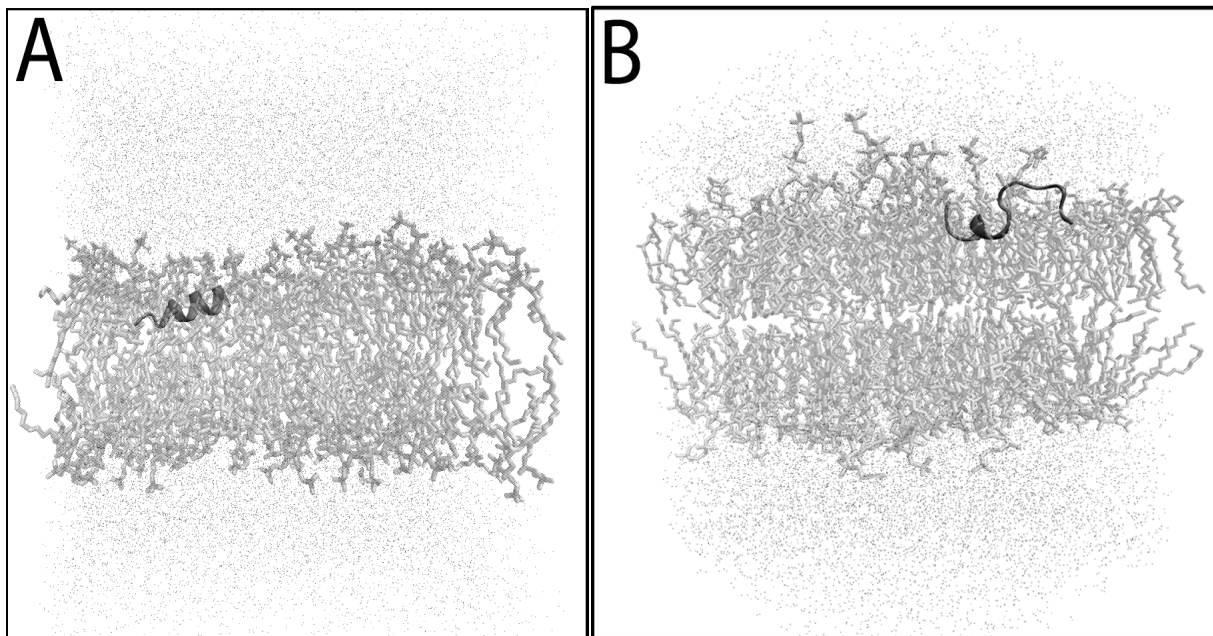


Figure 2: **A)** DMPC lipid bilayer showing 14-residue C-terminal MBP peptide after an 80 ns simulation. Note that the N-terminus of the peptide has deeply penetrated at least halfway into the leaflet. The helical structure of the peptide is preserved, compared to the initial state; **B)** DMPC/DMPE mixed bilayer showing 14-residue C-terminal MBP peptide. Observe that the helical structure has been greatly compromised after 80 ns. The peptide penetrated the DMPC/DMPE bilayer less significantly deeper compared to the DMPC bilayer possibly due to electrostatic interactions and a decrease in steric hindrance (see **Fig. 2A**).

3.3 Membrane penetration differences by MBP C-terminal peptide – Hypothesis

Synthetic DMPC and DMPE molecules represented phosphocholine (PC) and phosphoethanolamine (PE). Choline has three CH_3 groups attached to an N atom, while ethanolamine has three H atoms (see **Fig. 3**). Both choline and ethanolamine have the same formal charge of +1, but behave differently [3]. Nevertheless, substitution of ethanolamine for choline in the bacterial cell wall significantly alters important biological functions, such as cellular adhesion and bacterial transformation [9]. Zull and Hopfinger [3] measured the

accessibility of a negative test charge to a positively-charged N atom, concluding that ethanolamine interacts more strongly with anions. Even though choline lipids have a three times stronger partial positive charge on an N atom, the positive formal charge is sterically poorly accessible, which could result in poorer interactions with anions. The positive charge of ethanolamine on an N atom is more diffused and more accessible [3]. The negative N-terminus of the peptide could be thought of as a negative charge. Indeed, our results indicate that the peptide interacted more strongly with DMPE lipids in the mixed DMPC/DMPE membrane, as compared to the pure DMPC membrane, probably due to differences in the N positive charge accessibility. Thus, the negatively-charged N-terminus of the peptide stayed attached to the surface of DMPC/DMPE membrane, not being able to penetrate the membrane further. The opposite was observed in simulations with a DMPC bilayer. The partially shielded positive N charge of the DMPC bilayer was not as effective in capturing the negative N-terminus of the peptide, resulting in deeper penetration into the bilayer.

3.4 Comparison of our results to external evidence – Significance

The above results highlight the importance of the membrane composition, in conjunction with an array of other membrane properties, on final protein structural stability and behavior that is ultimately reflected in its biological function. These findings further support evidence from other studies that protein-membrane interactions and α -helical protein stability are governed by combination of factors including: hydrogen bonds, ion pairs, favorable surface van der Waals interactions, and thermodynamic parameters [12]. We were able to confirm that the interactions of the individual amino acids of a peptide with each other and with the surrounding medium (e.g., membrane lipids and polar water) determine the peptide's final structure, stability, and interaction behavior [13].

4 CONCLUSIONS

This study shows the usefulness of computational MD approaches in studying the conformations of biological membranes, particularly the effects of various factors affecting their fluidity and protein stability. The DMPC membrane showed a greater degree of fluidity at 309K compared to the mixed DMPC/DMPE membrane, with diffusion coefficients of $0.0207 \times 10^{-5} \text{ cm}^2/\text{s}$ and $0.0068 \times 10^{-5} \text{ cm}^2/\text{s}$. The DMPC membrane was more strongly penetrated by a 14-residue α -helical MBP C-terminal peptide compared to the mixed DMPC/DMPE bilayer. The peptide showed a greater stability retaining more of its α -helical structure in the DMPC membrane system compared to the mixed DMPC/DMPE one. These findings highlight possible dependence of MBP structure on membrane and sequence compositions, providing further insights into myelin structure. Enrichment of in DMPE molecules caused localized denaturation / instability of MBP C-terminal peptide α -helix. This finding suggests localized denaturation of solvent accessible MBP α -helices could possibly augment the degradation of myelin in MS, resulting eventually in subsequent decrease of nerve impulse propagation efficiency.

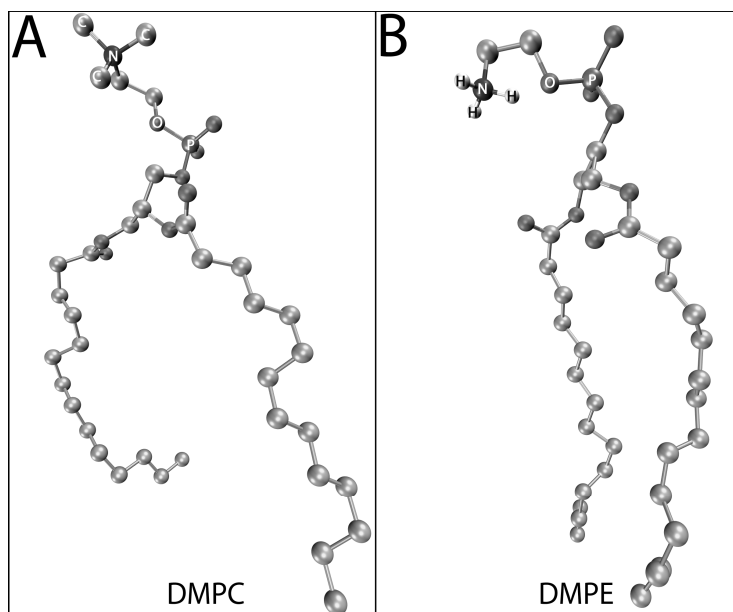


Figure 3: A) DMPC to B) DMPE lipid structural comparison explaining differences in accessibility to formal positive charge of the N atom. Some important atoms around the N atom are labeled with white letters.

ACKNOWLEDGMENTS

This work was supported by the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada. The author is grateful to Dr. George Harauz of the University of Guelph for his support.

References

- [1] Kuchel W., Philip W. and Gregory B. Ralson, *Theory and Problems of Biochemistry*, New-York, Schaum's Outline Series, McGraw-Hill, 1988.
- [2] Spector A. A., Yorek M. A., "Membrane lipid composition and cellular function," *J Lipid Res* **26**, pp. 1015-1035, 1985.
- [3] Zull, J. E., Hopfinger A. J., ". Potential energy fields about nitrogen in choline and ethanolamine: biological function at cellular surfaces," *Science* **165**, pp. 512-513, 1969.
- [4] Kandt C., Ash, W. L., Tieleman D.P., "Setting up and running molecular dynamics simulations of membrane proteins," *Methods* **41**, pp. 475-488, 2007.
- [5] Martin C. E., Hiramitsu K., Kitajima Y., Nozawa Y., Skriver L., Thompson G. A., "Molecular control of membrane properties during temperature acclimation. Fatty acid desaturase regulation of membrane fluidity in acclimating Tetrahymena cells," *Biochemistry* **15**, pp. 5218-5227, 1976.
- [6] Jeschke M. G., Klein D., "Liposomal gene transfer of multiple genes is more effective than gene transfer of a single gene," *Gene Ther* **11**, pp. 847-855, 2004.
- [7] Harauz G., Ladizhansky V., Boggs J. M., "Structural polymorphism and multifunctionality of myelin basic protein," *Biochemistry* **48**, pp. 8094-8104, 2009.
- [8] Harauz G., Musse A. A., "A tale of two citrullines--structural and functional aspects of myelin basic protein deimination in health and disease," *Neurochem Res* **32**, pp. 137-158, 2007.
- [9] Tomasz A., "Choline in the cell wall of a bacterium: novel type of polymer-linked choline in *Pneumococcus*," *Science* **157**, pp. 694-697, 1967.
- [10] E. Lindahl et al., "GROMACS 3.0: a package for molecular simulation and trajectory analysis," *J. Mol. Model.* **7**, pp. 306-317, 2001.
- [11] Hopp T.P., Woods K.R. "Prediction of protein antigenic determinants from amino acid sequences" *Proc. Natl. Acad. Sci.* **78**, pp. 3824-3828, 1981.
- [12] Popot JL, Engelman DM., "Helical membrane protein folding, stability, and evolution" *Annu Rev Biochem* **69**, pp. 881-922, 2000.
- [13] William C., "Annual Review of Biophysics and Biomolecular Structure", *Annu Rev Biochem* **28**, pp. 319-365, 1999.

The application of Quantum Tunneling Compound to sleep actigraphy

Martin Bowyer^a and Ken Jones^a and Mila Kwiatkowska^a

^aThompson Rivers University, 900 McGill Road, Kamloops, Canada

ABSTRACT

The purpose of this paper is to describe the application of material constructed from Quantum Tunneling Compound (QTC) to the problem of detecting and recording movement during sleep. We describe the design and implementation of a pressure sensitive mat (PSM) incorporating QTC technology. Furthermore, we describe the neural fuzzy analysis of actigraphic data.

Keywords: Actigraphy, Polysomnography, QTC, Sleep

1 INTRODUCTION

This paper proposes the design of a Pressure Sensitive Mat (PSM) based on Quantum Tunneling Compound (QTC) for the collection of actigraphic data during a sleep study. The incorporation of QTC technology provides a higher level of precision and accuracy when compared with accelerometer or strain sensor technologies. QTC is a material which can be easily integrated into electronic circuits in order to provide pressure sensing capabilities. An increase in the precision and accuracy of detection and recording of movement during sleep is important since such measurements increase the value of sleep studies and thus cause a decrease in false negatives when testing for sleep disorders. The PSM system would support the diagnosis of three specific sleep disorders which include Obstructive Sleep Apnea Syndrome (OSAS), Periodic Limb Movement Syndrome (PLMS), and Restless Leg Syndrome (RLS).

The term 'sleep apnea' refers to a sleep disorder in which "oxygen desaturation and carbon dioxide retention" occur during sleep, activating the sympathetic nervous system and disturbing sleep[1]. The resulting sleep disturbance leads to many health problems such as daytime hypertension[1], anxiety, depression, structural alterations in the brain[2] as well as safety problems such as increased risk of vehicle accidents[3].

PLMS refers to a sleep disorder in which the patient is unable to resist limb movement when at rest. The uncontrolled limb movement makes it difficult to get a good nights sleep and increases the probability of developing insomnia, anxiety, and depression[4]. Similar to PLMS, RLS is a sleep disorder in which the patient is unable to resist random leg movements at all times.

Actigraphy is defined as the measurement of movement. Most actigraphs in use today are wrist actigraphs[5]. Additionally, bed actigraphy[6] has been validated as a non-constraining actigraphic method. Wrist actigraphy is performed using an accelerometer based, watch-like device[5] which is attached to the patients wrist or ankle. Wrist actigraphs are mildly constraining and patients, especially infants, may find it difficult to sleep with the device. Bed actigraphy involves placing the bed on strain pressure sensors and calculating changes in bed pressure over time. While non constraining, current bed actigraphy methods only identify the difference between wake and sleep[6] whereas the proposed PSM system will identify positioning on the bed in addition to detecting and recording movement.

2 PSM Description

This section describes QTC material and its implementation for PSM in sleep actigraphy.

2.1 QTC

QTC is a special type of electrically conductive material with pressure sensing capabilities. The electrically conductive material is composed of nano-sized metal particles with irregular spiked surfaces insulated by a silicone rubber[7]. This combination creates a special material that becomes more conductive when pressure is applied.

The term quantum tunneling originates from the fact that electrons can in some cases be described as waves and that there is a finite probability that an electron is able to tunnel through a normally forbidden barrier. In the case of QTC, electrons must tunnel through the isolating silicone rubber in order to reach a neighboring conductive particle. In practice, QTC is used as a variable resistor between two potential voltages. By measuring the impedance

of the circuit it is possible to determine how much force is being applied to the material. Also, because there is no activating pressure required, the full voltage range is available to be utilized.

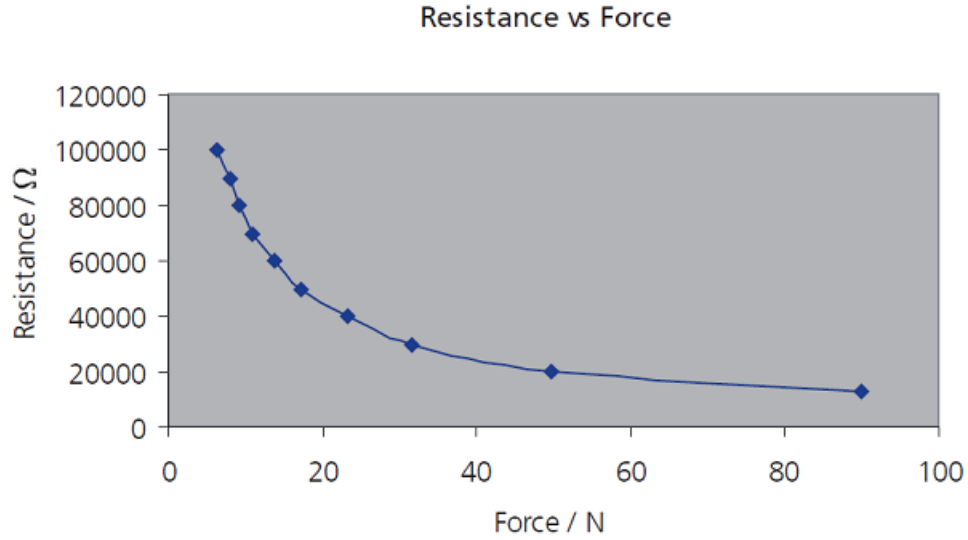


Figure 1. The impedance caused by applying force[8]

2.2 Implementation of QTC

The construction of the QTC sheet consists of three layers: a charged conductive layer, the QTC material and a wire grid. A voltage defined as V_{in} is applied to the charged conductive layer and HIGH-Z is applied to the wire grid. A weighted graph can then be created by analyzing the voltage at intersection points across the wire grid.

The wire grid allows the measure of the voltage level at a point defined by the intersection of vertical and horizontal wires. This is accomplished using switches on both the vertical and horizontal wires in order to allow only 1 circuit pathway at a time. This pathway passes through an op-amp circuit which produces an output voltage inversely proportional to the resistance[9] as represented by the following equation:

$$V_{out} = -V_T \times \frac{R_F}{R_S} \quad (1)$$

The advantage of using an op-amp circuit is that it produces a linear change in the output voltage. An analog to digital converter uses V_{ref} in order to output an 8bit value describing V_{out} in relation to V_{ref} at the intersection point. Figure 2 illustrates an example of an op-amp circuit where R_S represents the resistance from the QTC material.

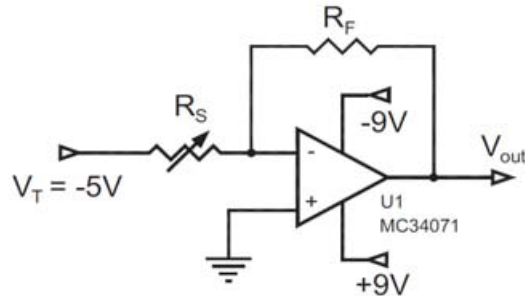


Figure 2. Example circuit illustrating how QTC is used as a variable resistor affecting the output signal[9]

The PSM circuit consists of gates to control which wire intersections are active, an analog to digital converter in order to digitize the measurement as well as a control unit and communication port to transmit the the data. Using a wire grid allows the system to iteratively take measurements for each row and column intersection. The data output is a contiguous block of bytes with a size of rows \times columns \times 1 byte. When interpreting the data, each row is

represented by a stride of data. A stride of data is the equivalent byte size of columns \times 1 byte. The total number of strides is equivalent to the total number of rows. The construction of the PSM consists of a QTC sheet over-top of a dense rubber layer. The purpose of the dense rubber layer is to impede the dissipation of the force being applied to the PSM. Impeding the dissipation of the force is important in order to produce accurate measurements of force. Figure 3 illustrates a cross-section of the PSM. Furthermore, the QTC material is calibrated by adjusting the density of the nano-particles. Selecting a nano-particle density from the lower end of the scale allows the overall thickness of the QTC material to be reduced without affecting the impedance property.

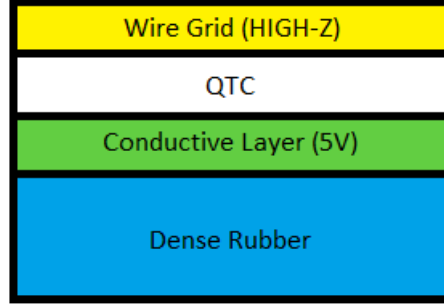


Figure 3. Cross-section of PSM construction

3 Data Analysis

There are two questions which we wish to answer using the actigraphic data output from the PSM; The first question is 'What is the current position of the patient laying on the PSM?', the second question is 'What change has been made in the position of the patient over time?'. The approach we take in answering these questions incorporates the use of artificial neural networks (ANN) and fuzzy sets (FS). In order to more easily analyze the data, the stored output from the PSM is first decoded into a set of 3 dimensional tuples $\vec{p}_i = \langle x, y, z \rangle$ where $i \in (1, n)$ where $n \in \mathbb{N}$ is the number of sensor points embedded in the PSM. The first two coordinates x, y identify the location on the planar sheet while the third coordinate z , identifies a measure of the pressure being applied at that location. These coordinates partition the PSM into a set of rectangles bordered by 4 sensor points. Our data analysis begins with associating one input node of an ANN with each sensor point. We will be using a three layer ANN with n nodes in the input layer, $n - 2$ nodes in the hidden layer and n nodes in the output layer. So, $in_i(t) = \vec{p}_i = \langle x, y, z \rangle$ for the i^{th} input node. The input value for each hidden layer node coming from the input layer is calculated as:

$$net_j(t) = \sum_{i=1}^n w_{ij}(t) o_i(t) \quad (2)$$

Where $net_j(t)$ is the weighted sum of all inputs to hidden node j , $j \in [1, n - 1]$. w_{ij} represents the weight associated with the edge joining input node i to hidden node j and $o_i(t)$ is the value output from input node i .

The activation function we're using for the hidden layer is the sigmoid[10]:

$$f_j(net_j(t)) = \frac{1}{1 + e^{-net_j(t)}} \quad (3)$$

Since we wish to identify numerous positions and types of movement, a different ANN will be used for each with the same construction. As we do not currently have a sample of the QTC material, these training sets are generated using simulation. The training of each network is accomplished using the standard backpropagation technique using an error threshold of 0.001.

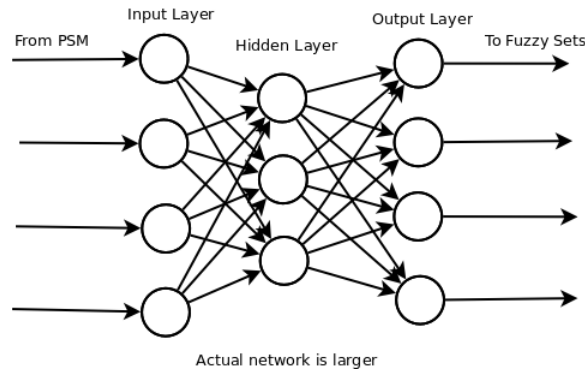


Figure 4. Triple Layer ANN

In order to identify the presence of movement and its type, data samples are tested from time t to time $t + q$ and their identification as per the neural networks they match is counted. The relative frequency of the matching for each network is calculated in order to identify movement. If no movement is identified, both questions have been answered, if movement is identified we continue. Each type of movement we are interested in is associated with a fuzzy set, the difference between samples is calculated for each node in order to create a set of output values which is then scored as before. Using the maximum type frequency as the membership function, we identify the type of motion. With this method we have answered both of the questions above.

4 CONCLUSION

In this paper we have described the design of an actigraphic method. This method has the potential to increase the usefulness of actigraphy in sleep studies for the diagnosis of OSAS, RLS, and PLMS. PSM is most useful in situations where standard methods of actigraphy are not sufficient. Situational usefulness of this method includes when a patient has skin hypersensitivity, or when precise measuring of body positioning is important. Additional medical applications of QTC PSM include gait analysis, physiotherapy, and posture analysis. The validation of the QTC method should be a subject of future study.

References

- [1] L. M. Prisant, T. A. Dillard, and A. R. Blanchard, "Obstructive sleep apnea syndrome.," *Journal Of Clinical Hypertension (Greenwich, Conn.)* **8**(10), pp. 746 – 750, 2006.
- [2] R. Kumar, P. M. Macey, R. L. Cross, M. A. Woo, F. L. Yan-Go, and R. M. Harper, "Neural alterations associated with anxiety symptoms in obstructive sleep apnea syndrome.," *Depression And Anxiety* **26**(5), pp. 480 – 491, 2009.
- [3] Y. Komada, Y. Nishida, K. Namba, T. Abe, S. Tsuiki, and Y. Inoue, "Elevated risk of motor vehicle accident for male drivers with obstructive sleep apnea syndrome in the tokyo metropolitan area.," *The Tohoku Journal Of Experimental Medicine* **219**(1), pp. 11 – 16, 2009.
- [4] S. Sevim, O. Dogu, H. Kaleagasi, M. Aral, O. Metin, and H. Camdeviren, "Correlation of anxiety and depression symptoms in patients with restless legs syndrome: a population based survey.," *Journal Of Neurology, Neurosurgery, And Psychiatry* **75**(2), pp. 226 – 230, 2004.
- [5] M. P. Rothney, G. A. Apker, Y. Song, and K. Y. Chen, "Comparing the performance of three generations of actigraph accelerometers.," *Journal Of Applied Physiology (Bethesda, Md.: 1985)* **105**(4), pp. 1091 – 1097, 2008.
- [6] B. H. Choi, J. W. Seo, J. M. Choi, H. B. Shin, J. Y. Lee, D. U. Jeong, and K. S. Park, "Non-constraining sleep/wake monitoring system using bed actigraphy.," *Medical & Biological Engineering & Computing* **45**(1), pp. 107 – 114, 2007.
- [7] D. Bloor, A. Graham, E. J. Williams, P. J. Laughlin, and D. Lussey, "Metal–polymer composite with nanostructured filler particles and amplified physical properties," *Applied Physics Letters* **88**(10), p. 102103, 2006.
- [8] Peratech, "Qtc force sensors," January 2004.

- [9] C. Doggen, “Floor sensor for 3tu humanoid,” December 2009.
- [10] D. M. Skapura, *Building Neural Networks*, acm press, 2002.

A Fast Technique of Tissue Biomechanical Analysis for Real-time Prostate Tissue Elasticity Reconstruction

S. Reza Mousavi^{a,1} and Abbas Samani^{a,b,c}

^aDepartment of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada;

^bDepartment of Medical Biophysics, The University of Western Ontario, London, ON, Canada;

^cImaging Research Laboratories, Robarts Research Institute (RRI), London, ON, Canada

ABSTRACT

Elastography image reconstruction techniques typically involve displacement or stress field calculation of tissue undergoing mechanical stimulation that can be done by Finite Element (FE) analysis. However, traditional FE method is time-consuming, and hence not suitable for real-time or near real-time applications. In this article, we present an alternate accelerated method of stress calculation that can be incorporated in elastography reconstruction algorithms. Shape is an essential input of FE models that is considered in conjunction with material stiffness and loading to yield stress distribution. The essence of the proposed technique is finding a function between shape and stress field. This function takes the shape parameters as input and outputs the stress field very fast. To develop such a function principal component analysis (PCA) is used to obtain the main modes of shape and stress fields. As such, the shape and stress fields can be described by these main modes weighted by a small number of weight factors. Then, an efficient mapping technique is developed to relate the weight factors of shape to those of the stress fields. We used Neural Network (NN) for this mapping, which is the sought function required to input shape and output stress field. Once the mapping function is obtained it can be used for analyzing shapes not included in the NN training database. We employed this technique for prostate tissue stress analysis. For a typical prostate, our results indicate that analysis using our technique takes less than 0.07 seconds on a regular desktop computer irrespective of the model size and complexity. This analysis indicates that stress error of the majority of the samples is less than 5% per node.

Keywords: Prostate Cancer, Stress Analysis, Finite Element, Real-time, Principal Component Analysis

1 INTRODUCTION

Prostate cancer is the most common cancer in Canadian men. Prostate tumors usually grow slowly, and if detected early, it can often be cured or managed successfully [1]. For many years Digital Rectal Examination (DRE), Prostate-Specific Antigen (PSA) and Trans Rectal Ultra Sonography (TRUS) have been the primary techniques for prostate cancer detection [2]. However, these conventional methods have low sensitivity and specificity for prostate cancer detection [3]. For instance, comparison of TRUS-based diagnosis of prostate cancer to pathological evaluation (gold standard) found that ultrasound based diagnosis has a sensitivity of 52% and a specificity of 68% [4]. In contrast, it has been shown that there is a strong correlation between pathological and mechanical properties of soft tissue [5]. As such, based on the fact that variations in tissue elastic properties are associated with the presence of cancer [6], elastography in conjunction with US imaging can detect prostate cancer with a higher sensitivity [7].

Ultrasound elastography is a novel imaging technique in which elastic properties of tissues are reconstructed and displayed. Elastography image reconstruction techniques typically involve displacement or stress field calculation of tissue undergoing mechanical stimulation. This can be done by Finite Element Method (FEM), which is time-consuming, hence is not suitable for real-time or near real-time applications. In this work, we present an alternate accelerated method of tissue stress calculation that can be incorporated in real-time elastography reconstruction algorithms. This method develops a mapping scheme between shape space (e.g. different prostate shapes) and stress space. This mapping function can calculate the tissue stress field in real-time or near real-time fashion.

¹ Corresponding author. E-mail: smousav8@uwo.ca, Telephone: +1(226)374-4776.

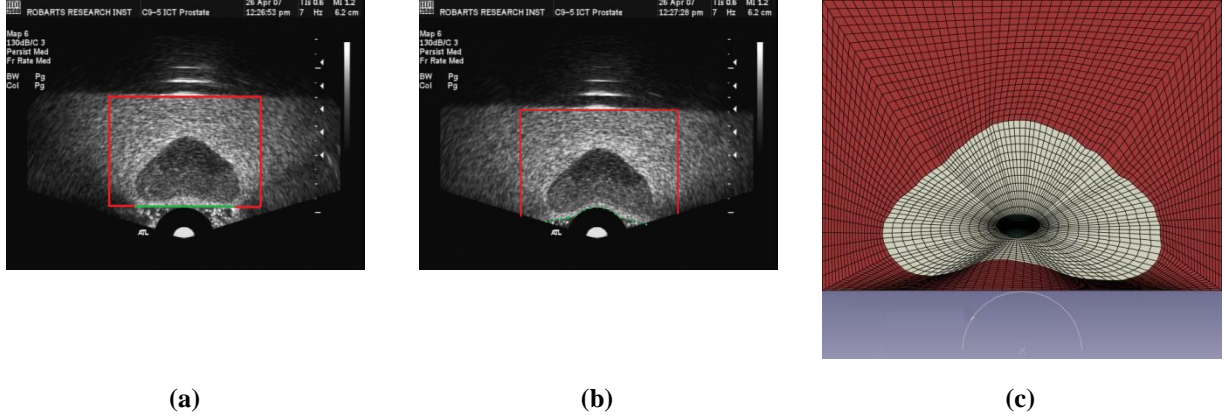


Figure 1. 2D TRUS Images (a) before compression and (b) after compression. Points on the green line are free to move while points on the red lines are almost fixed. (c) Sample prostate-tumor model.

2 METHODOLOGY

2.1 Modeling

In order to calculate a displacement or stress field of tissue undergoing mechanical stimulation, FEM modeling can be used which requires the geometry and biomechanical properties of the tissue and boundary conditions. In this work, 2D TRUS images were used to construct the model. Based on these images (Fig. 1), mechanical stimulation is applied to the bottom of the prostate using ultrasound probe, which compresses the prostate and its surrounding tissue. The prostate tissue along with a block of surrounding tissue is incorporated in the model since the effect of the probe compression becomes insignificant at points far away from its application region. Hence, as shown in Fig. 1c, our model contains the prostate with a tumor inside a rectangular area mimicking the surrounding connective tissue. All points on the rectangle's edges are fixed except some points in the middle of the bottom edge where the probe applies compression. Different Young's moduli were assigned to the three regions of the tumor, prostate and surrounding tissue, and the model is discretized into a FE mesh. As the load acts in the plane of the 2D model (with small thickness) the problem is idealized as a plane stress problem.

2.2 FE Mapping Function

FEM is a time-consuming method; therefore, it is not suitable for real-time elasticity reconstruction. Tissue stress or displacement calculation can be accelerated if FEM is substituted by a mapping function that maps prostate shape into displacement and stress fields for a given loading. Establishing such a mapping function is possible because inter-patient prostate shape variability is modest while tissue deformation and stress distribution patterns under a given clinical mechanical stimulation are expected to be similar. Each prostate-tumour configuration can be represented by a set of points located on the boundary of the prostate and on the boundary of the tumour, called "landmarks". In order to compare equivalent points from different shapes, all shapes are aligned by scaling, rotation and translation with respect to a set of axes. Considering n landmarks, each shape in the shape space is given as follows:

$$X_i = \{x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2}, \dots, x_{i,n}, y_{i,n}\} \quad i = 1, \dots, N \quad (1)$$

where (x_i, y_i) are the coordinates of each landmark and N is the number of shapes in the shape space. As discussed earlier, each model should be meshed to be suitable for FE analysis. In this work, each shape was discretized using a common TFI-based FE mesh with quadrilateral elements [8], resulting in m elements for each shape. Conventional FE analysis provides accurate stresses at the elements' centroids [9]. Hence, different stress fields of each shape (e.g. σ_{yy}) obtained from FE analysis can be given as follows:

$$S_{yy,i} = \{s_{yy,i1}, s_{yy,i2}, \dots, s_{yy,im}\} \quad i = 1, \dots, N \quad (2)$$

Because of the large array size of X and S , it is not efficient to establish a mapping function directly between vectors of $2n$ -D shape space and their corresponding vectors in m -D stress space. This may result in a complicated mapping function with a large number of parameters to be tuned. In order to have an efficient mapping, we find the main modes or principal components of both shape space and stress space, and then map the weight vectors of each space to their stress weight vectors counterparts.

2.2.1 Principal Component Analysis (PCA)

In PCA, main modes are specified by calculating the eigenvectors and eigenvalues of the covariance matrix of a space. Considering a space with N points, the covariance matrix of it is defined as:

$$Cov = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T, \quad \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (3)$$

Eigenvectors of Cov are the orthogonal components of this space and their corresponding eigenvalues show how significant they are. The larger the eigenvalue the more significant is the corresponding eigenvector. Hence, based on the eigenvalues of the shape space, the L most significant eigenvectors $P = (p_1 p_2 \dots p_L)$ are adopted as the main modes of the shape space such that the ratio of the sum of the corresponding L eigenvalues to the sum of all eigenvalues is more than 0.99. Similarly, the T most significant eigenvectors of the stress space $Q = (q_1 q_2 \dots q_T)$ are adopted as the main modes of that space. According to PCA, the vectors of each space are mapped to its main modes resulting in vectors of weight factors:

$$\begin{aligned} X_i &= \bar{X} + b_i P \Rightarrow b_i = P^+ (X_i - \bar{X}) \\ S_i &= \bar{S} + c_i Q \Rightarrow c_i = Q^+ (S_i - \bar{S}) \end{aligned} \quad (4)$$

In which P^+ and Q^+ are the pseudo-inverse matrices of P and Q , respectively.

Stress field of tissue undergoing mechanical stimulation depends on both shape and Young's modulus distribution. Therefore, the Young's moduli of tissues are added to the weight factors of points in the shape space and the mapping function is established between the resulting augmented vectors ($[b_i E_i]$) and vectors of weight factors in the stress space (c_i).

2.2.2 Mapping Function Computation

We use Neural Networks (NN) to relate shapes and stress fields. The NN we used for this purpose is a multi-layer feed-forward back propagation neural network. In general, Multilayer Feed Forward Neural Network (FFNN) [10] is widely used in function approximation applications. Such networks consist of an input layer, which conducts the inputs to the next layer, a number of hidden layers and an output layer. Hidden and output layers include a number of neurons. Each neuron receives a number of weighted inputs as well as a bias and yields an output. To compute its output, each neuron uses a transfer function over the sum of its weighted inputs and bias. During the training phase, the network finds an optimum mapping relationship between the input and output vectors using training samples, i.e. a number of input vectors and their corresponding known output vectors. This is carried out by the network through adjusting its neurons' weights and bias values to minimize the differences between the network's known responses to their respective input samples. The most common training algorithm used in FFNN is the back-propagation algorithm, which is based on the gradient descent method. The term back-propagation refers to the manner in which the gradient is computed for nonlinear multilayer networks. In the simulation phase, the trained network responds to new input vectors based on its knowledge achieved during the training phase to produce the output. In this study, a three-layer feed-forward back-propagation neural network was applied for function approximation. The NN's topology was chosen such that the input layer has the size of input vector $[b_i E_i]$ with one hidden layer consisting of 15 neurons in addition to the output layer. The output layer includes as many neurons as the size of c_i . All the neurons used 'tansig' as their transfer function except the output neurons that used 'purelin' as transfer function.

3 RESULTS

A database of 1000 prostate-tumour configurations was produced to evaluate the proposed method. The fitting NN was trained with 800 out of the 1000 samples. 200 additional samples were then used to test the mapping function. Results were validated by FE analysis results obtained by ABAQUS (commercial FEM software). Figure 2 shows the average error per node for 200 test samples. Figure 2 indicates that the majority of the samples have errors of less than 5% while only very few samples encounter errors larger than 10%. The latter ones are the ones that correspond to very small stress values, hence their percentage error is amplified. Stress fields resulting from conventional FE analysis and from FFNN function for a typical test sample are depicted in Fig.3. This figure shows a very good agreement between these fields. Figure 4 shows the difference of the two result sets. It indicates that the difference in regions near the contact nodes and boundaries of prostate and tumor is higher than other regions.

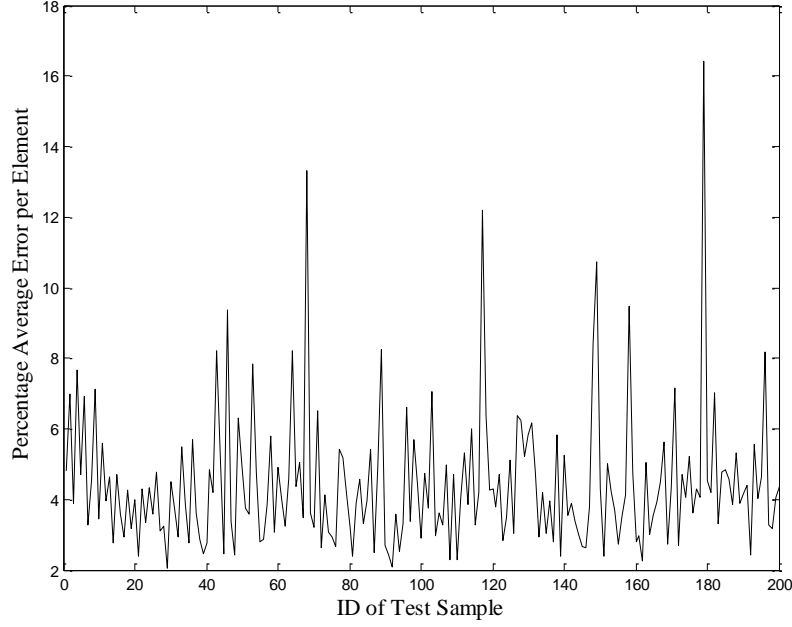


Figure 2. Percentage average error per element of stress field for 200 test samples (using 800 training samples).

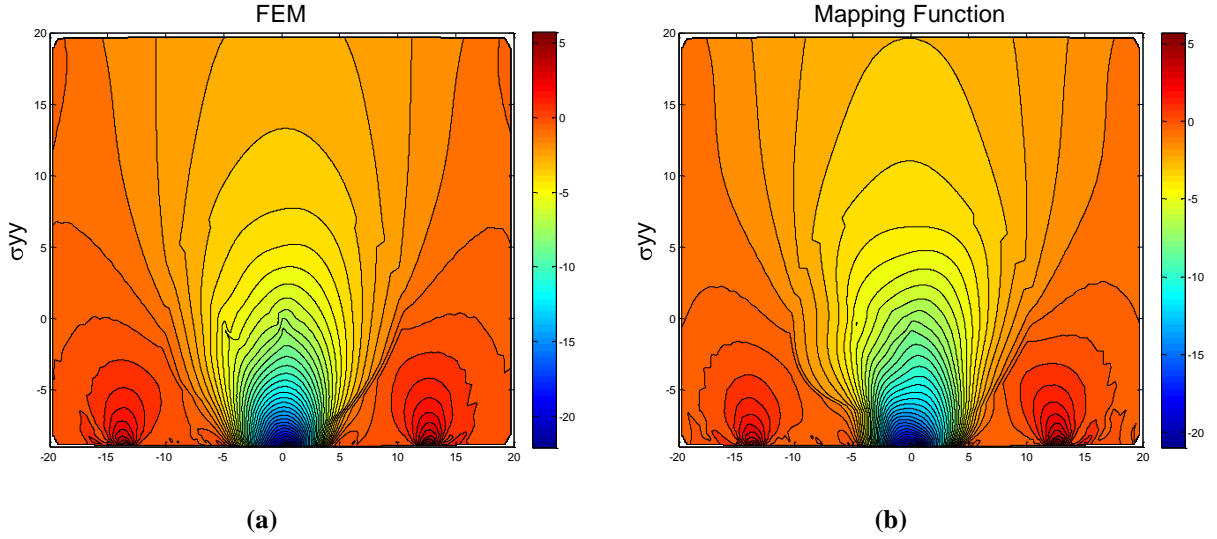


Figure 3. Stress field resulted from (a) FEM and (b) FFNN mapping function.

4 CONCLUSION AND DISCUSSION

In this paper, we presented a fast method for estimating stress field of tissue under specified loading conditions. The proposed method establishes a mapping function to relate shape space and stress space. Due to the large number of variables required to define the shape and stress spaces, PCA was employed to reduce the dimensions by projecting both the shape and stress spaces to their main modes. The resulting compact spaces were then interrelated via a neural network model. The proposed method is both fast and accurate for calculating stress field of the same class of objects. Further work is under way to use this mapping for our new real-time elastography modulus reconstruction technique in which prostate and tumor moduli are updated iteratively using strain images acquired from an ultrasound imaging system and stress field estimated with the proposed method.

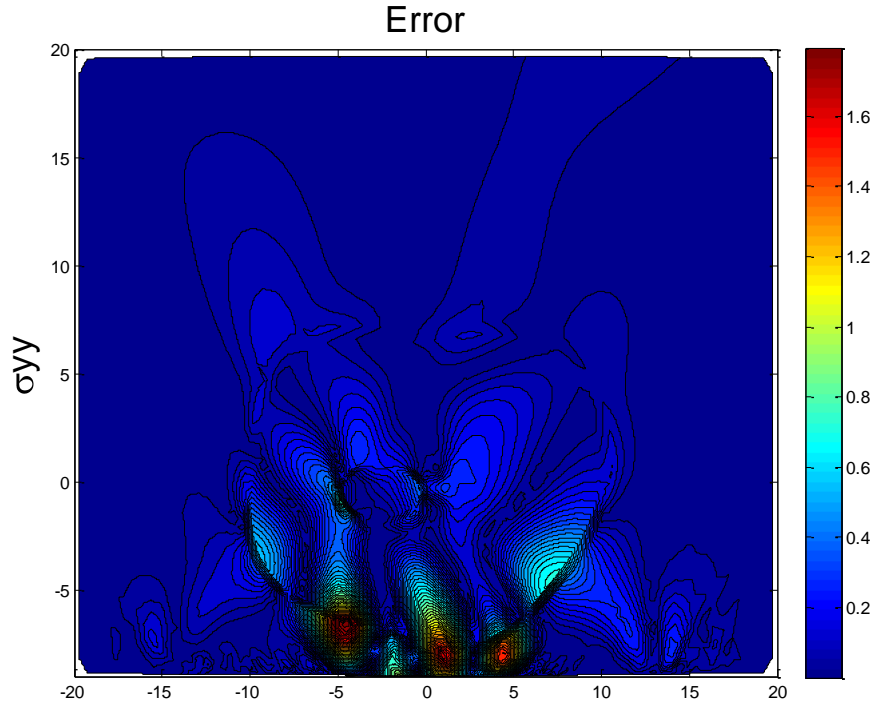


Figure 4. Difference of Stress fields resulting from FEM and FFNN mapping function.

References

- [1] Canadian Cancer society: www.cancer.ca.
- [2] K. Kamoi, K. Okihara, A. Ochiai, O. Ukimura, Y. mizutani, A. Kawauchi, and T. Miki, "The Utility of Transrectal Real-Time Elastography in the Diagnosis of Prostate Cancer," *Ultrasound in Med. & Biol.* **34**, pp. 1025–1032, 2008.
- [3] G. Salomon, J. Kollerman, I. Thederan, F. K.H. Chun, L. Budaus, T. Schlomm, H. Isbarn, H. Heinzer, H. Huland, and M. Graefen, "Evaluation of Prostate Cancer Detection with Ultrasound Real-Time Elastography: A Comparison with Step Section Pathological Analysis after Radical Prostatectomy," *European Urology* **54**, pp. 1354–1362, 2008.
- [4] H. B. Carter, U. M. Hamper, and S. Sheth, "Evaluation of Transrectal Ultrasound in the Early Detection of Prostate Cancer," *J. of Urol.* **142**, pp. 1008-101, 1989.
- [5] T. A. Krouskop, T. M. Wheeler, F. Kallel, B. S. Garra, and T. Hall, "Elastic Moduli of Breast and Prostate Tissues under Compression," *J. of Ultrasound Imaging* **20**, pp. 260-274, 1998.
- [6] A. Samani, J. Bishop, and D. B. Plewes, "A Constrained Modulus Reconstruction Technique for Breast Cancer Assessment," *IEEE Trans. Medical Imaging* **20**, 2001.
- [7] K. Konig, U. Scheipers, A. Pesavento, A. Lorenz, H. Ermert, and T. Senge, "Initial Experiences with Real-Time Elastography Guided Biopsies of the Prostate," *J. of Urol.* **174**, pp. 115-117, 2005.
- [8] P. M. Knupp, S. Steinberg, *Fundamentals of Grid Generation*, CRC Press, 1994.
- [9] *ABAQUS/Standard User's Manual*, Habbitt, Kalson & Sorensen, Inc., Version 6.3, 2002.
- [10] N. Toda, K. I. Funahashi, S. Usui, "Polynomial functions can be realized by finite size multilayer feedforward neural networks," *Proc. IEEE Int. Joint Conf. Neural Networks*, pp. 343-384, Singapore. 1991.

Finding Important Protein Motions in Solution Through Linear Dimensionality Reduction

Krzysztof Borowski* and Forbes Burkowski

University of Waterloo, Waterloo, Canada

ABSTRACT

Automating protein structure analysis is useful to understand their function, especially as structure data grows. Developing methods of elucidating protein motions from the multitude of data currently available, and doing it in a way that allows non-experts to easily perceive protein movement can lead to heightened levels of understanding and hypothesis generation. We apply principal component analysis to examine major modes of motion for protein conformations. We show that a few principal components of the conformation matrix can capture the majority of the motions, and present a visual depiction of one mode of calmodulin.

Keywords: protein flexibility, PCA, SVD, NMR, calmodulin

1 BACKGROUND

As datasets from wet-lab experiments on protein structure grow in size and number, the interpretation of these results becomes difficult. X-ray Crystallography experiments are able to produce a large amount of data about crystallized protein structure[1]. For example, there are over 150 conformations of the protein HIV-1 Protease available on the Internet. Due to its pharmaceutical importance (as a drug target in HIV research), it is a valuable endeavor to understand the main modes of movement of this protein. Principal Component Analysis (PCA) has been applied to HIV-1 Protease before, with results showing a similar mobility to movement determined experimentally. It has also been used in conjunction with singular value decomposition (SVD) to that same end with other protein sets[2, 3].

Nuclear Magnetic Resonance (NMR) is another method of attaining structural protein data[4]. Here, the structure of protein is taken from protein in solution, which allows for more variability in the data collected. Consequently NMR can create even more data than X-ray Crystallography. Many solution-NMR experiments have resulted in multiple models, each presenting a different conformation for the protein in question. With many such models within a file, and with many such files, it becomes imperative to develop automated methods of discerning major protein motions in a cost efficient way. PCA is a possible solution to this issue.

Other methods, such as normal mode analysis (NMA), are also applicable to finding protein motions. NMA considers harmonic motions and involves energy minimization. This makes NMA a more biologically relevant type of analysis when compared with PCA. However, NMA suffers from computationally demanding steps, and the results of PCA and NMA can be comparable[5].

In this paper, we implement an automated PCA based strategy for finding principal modes of motion of a protein set[2]. We develop Python code for use with UCSF Chimera[6] to present possible principal modes of motion for a given set of protein conformations, and we apply it to the protein calmodulin.

1.1 Principal Component Analysis

PCA is a dimensionality reduction technique which allows for high dimensional variables to be embedded into a low dimensional space where they are uncorrelated. These reductions in the lower dimension, known as principal components, are linear combinations of the original high dimensional variables. Each principal component is constructed to cover as much variance in the original data as possible; this allows for the mapping from high to low dimensional space to maintain information of the high dimensional system in very few principal components. It is possible to obtain as many principal components as there are original variables, but PCA is intended as a way of finding the smallest number of uncorrelated principal components which cover a large percentage of total variation in the original data. Even though this means some information is lost in the process, the trade-off between representing information through less variables may be beneficial in many applications. One such area is the study of molecular motions.

*Corresponding author. E-mail: kborowsk@uwaterloo.ca

1.2 Singular Value Decomposition

Singular Value Decomposition (SVD) is a common method in linear algebra in which a matrix is factored into a diagonal matrix, and two eigenvector matrices[7]. The SVD method allows the creation of a pseudoinverse for matrices which are close to being singular, but it also allows for extraction of important information from a matrix in a way similar to PCA[3]. It can be an efficient way of obtaining principal components. The SVD of an $m \times n$ matrix A is defined as:

$$A = UDV^T \quad (1)$$

where U and V are orthogonal left and right eigenvector sets of A respectively, and D is a nonnegative diagonal matrix with elements being the singular values of A . For our purposes, the trace of D is the total variance in A , while the square of each singular value represents the variance of the data in A along the corresponding vector in U [2].

2 METHODS: APPLYING PRINCIPAL COMPONENT ANALYSIS TO PROTEIN CONFORMATION

We will use SVD to gain the principal components of a set of molecular structures, thereby getting the principal modes of movement[2]. First, we acquire a set of l structures of the same protein. Each atom in the protein model has Cartesian coordinates (x_i, y_i, z_i) . Initially, we must complete a rigid least squares fit on all models with respect to one of the structures, which removes translational and rotational degrees of freedom[8].

We represent the entire protein model by creating a conformational vector of the form

$$m_i = (\Delta x_1, \Delta y_1, \Delta z_1, \Delta x_2, \Delta y_2, \Delta z_2, \dots, \Delta x_n, \Delta y_n, \Delta z_n) \quad (2)$$

which is a concatenation of the displacement of n atom coordinates from the arithmetic mean along the Cartesian axes[3]. Since the structures are superimposed after the least squares fit, this requires finding the average conformational vector and subtracting it from the conformational vector of the given protein model. Using all the atoms in the model is possible and may be desired for some experiments, but increases calculation time. We limit the conformation vectors to alpha carbons on each residue in the model only, which allows an examination of backbone motion specifically. Others have shown that the principal components are useful when considering atoms in binding sites alone[3].

Once each conformation vector m for all l models are acquired, the matrix A is created by column-wise concatenation of the l conformation vectors:

$$A = [m_1 | m_2 | \dots | m_l], A \in \mathbb{R}^{3n \times l} \quad (3)$$

where each m is the conformation vector representation of a model defined above.

After creating the matrix A , we proceed with the SVD of A as presented by Equation (1). The matrix of left eigenvectors U has columns u_i , and it is these columns which are the principal components of A . Each u_i shows the *mode* of motion of the atoms which were used in building the conformational vectors. One can think of u_i as a collection of direction vectors for n atoms, each representing their *mode* of motion at time point i of l [3]. Since this is a PCA approach, each column vector u_i will hold the general directions where most of the variability in atom position occurs. For example, the first three elements of u_1 would define the vector of the first principal component of motion of the first atom under consideration.

The right eigenvectors v_i found in V , are projections of A on the u_i vectors in the U matrix. As the A in our case is a protein conformation matrix, the elements of v_i are the locations of each atom along its principal component u_i . They can also be used to discern preferred protein conformations[3].

3 RESULTS OF APPLICATION TO CALMODULIN DATA

To test the methodology presented above, we applied the SVD based PCA to models of calmodulin created by Zhang et al. (1995)[9]. Calmodulin is a protein involved in a variety of cellular functions, including protein synthesis, gene regulation, cell motility, and cellular secretion[10]. It is composed of two main domains tethered with an alpha helix which allows for fluctuation in protein structure for target binding[11]. The large amount of displacement between conformations of calmodulin makes it an excellent example of the effectiveness of this method. The experimental data from Zhang et al. (1995) has 30 models discerned by NMR.

In Figure 1, we show the amount of variance each principal component exhibits on total variance in the original dataset. The largest singular value shows that the first principal component accounts of 29% of the total variance. To achieve 90%, 11 principal components must be taken into account.

In Figure 2, we visualize the mode described by the first principal component for each alpha carbon in calmodulin. The cylinders indicate the vector of the most important mode along which an atom will move, with the rounded end pointing in the direction of movement.

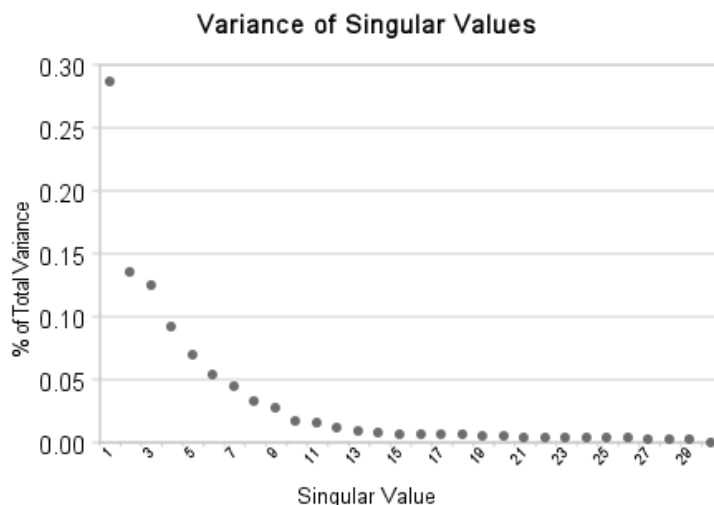


Figure 1.

Amount (%) of total variance of matrix A explained by specific singular values of A .

4 DISCUSSION

Previous work in this area has used the PCA to compare bound and unbound structures of ligand binding protein, and in the majority of cases, has been limited to data generated by X-ray Crystallography studies of proteins and their ligands[2]. Such research allowed for the principal modes of movement during binding to be understood. However, to our knowledge, NMR experiment data on unbound motions of a single protein have not yet been examined with this approach. The results suggest that the method is useful to discover transient motions in protein in solution by using multiple NMR models. By transient movement, we meant movement that is unrelated to binding specifically, but constitutes normal fluctuations in structure while a protein is in solution.

As Figure 1 shows, a majority of the modes of motion can be captured through a number of principal components much smaller than that of the models used in the analysis.

Even with the implied freedom of an unbound structure, the PCA method elucidated modes of movement similar to movement known as biologically relevant for calmodulin[11]. In Figure 2, we see the modes of one of the domains pointing in a direction where the folding movement may initially occur. Consecutive principal components, while not shown, suggest the same directions for the domain, while the 'anchored' domain shows very little mode activity at any point in time (according to the right eigenvectors). All of the atom modes seem to point in similar directions on the domain whose coordinate changes explain the movement of calmodulin. This is not visible in the 'anchored' domain: the principal components do not seem to point in similar directions, nor are they scaled by the right eigenvectors to the extent that the other domain is. This implies that it would be improbable for this domain to experience major movement in any particular direction, relative to the coordinate system of the original data. It is important to note that this is because the principal components are calculated relative to the protein structure coordinates of the original data: the data was pre-fitted around one of the domains, making it appear as though it is immobile to the PCA analysis. This anchoring provides a clearer view of the mobility of the other domain, but should not lead the reader to think the domain is somehow immobile in reality.

Overall, the principal motions of calmodulin in solution are based around the two domains coming closer and moving apart by the folding and extending activity of their alpha helix tether. Based on the data from Zhang et al. (1995)[9], this 'flopping' action occurs along a plane, which is why most of the modes seen in this work are close to being parallel directions (that is, if we were to translate them all to the axis along which this movement of calmodulin occurs, they would be as close to being on the plane itself).

5 CONCLUSION

Both the range of variability description found within the initial principal components, as well as the visualized modes show that the method can be used to better understand protein movement in solution. We used data on calmodulin in solution, whose 'domains on a tether' structure allows for various conformational possibilities in solution which fluctuate along a limited space of substates; they cannot behave without submitting physical and energy constraints. This is supported by the PCA analysis of the dataset, which suggest that major modes of movement

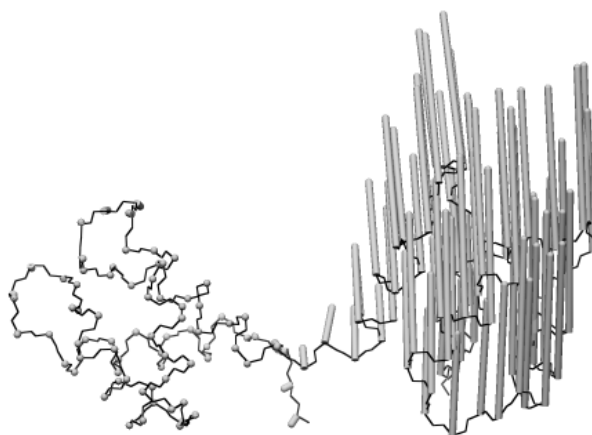


Figure 2.

The backbone of calmodulin with the first principal components, or modes of motion (left eigenvectors of A) indicated by the thin cylinders. The spherical end of the cylinder represents the positive direction of motion. The principal components are scaled by the right eigenvectors of A .

restrict motion of the protein along a specific path. We also show that the majority of said modes can be described with a few principal components found from the original data.

In the future, larger datasets should be used in order to find more modes. The limitation of structures derived from NMR experiments is that they are all found at the same experimental conditions. Combining multiple NMR experiment files to create input to PCA analysis may elucidate new modes of motion not found with a dataset from only one experiment. As well, this work only examined backbone movement; others have completed the PCA on all atoms, including those of residue sidechains[2]. Including sidechains in the analysis may shed light on more detailed movement of a protein in solution.

We used UCSF Chimera[6] to visualize the modes of calmodulin in a static fashion, but programs such as Chimera can be used to animate proteins undergoing changes along their principal components. Automated animation of structures about their main modes of movement would allow for quick analysis of probable protein mobility by non-experts.

Finally, PCA is a linear method of dimensionality reduction. Non-linear methods may improve results in dimensionality reduction applications. Analyzing main modes of motion may result in different outcomes if non-linear methods are used, such as locally linear embedding[12], Laplacian eigenmaps[13], or even kernel methods[14].

References

- [1] M. S. Smyth and J. H. J. Martin, "x ray crystallography," *Molecular Pathology* **53**(1), p. 8, 2000.
- [2] M. L. Teodoro, G. N. P. Jr, and L. E. Kavraki, "A dimensionality reduction approach to modeling protein flexibility," in *Proceedings of the sixth annual international conference on Computational biology*, p. 299308, 2002.
- [3] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. P. Jr, "Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements," *Proteins: Structure, Function, and Bioinformatics* **22**(4), p. 311321, 1995.
- [4] K. Wuthrich, "NMR of proteins and nucleic acids," *The George Fisher Baker non-resident lectureship in chemistry at Cornell University (USA)* **1**, 1986.
- [5] S. Hayward and B. de Groot, "Normal modes and essential dynamics," *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-* **443**, p. 89, 2008.

- [6] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF chimera visualization system for exploratory research and analysis," *Journal of computational chemistry* **25**(13), p. 16051612, 2004.
- [7] G. H. Golub and C. F. V. Loan, *Matrix computations*, Johns Hopkins Univ Pr, 1996.
- [8] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **34**(5), p. 827828, 1978.
- [9] M. Zhang, T. Tanaka, and M. Ikura, "Calcium-induced conformational transition revealed by the solution structure of apo calmodulin," *Nature Structural Biology* **2**, pp. 758–767, Sept. 1995. PMID: 7552747.
- [10] W. Y. Cheung, "Calmodulin plays a pivotal role in cellular regulation.," *Science (New York, NY)* **207**(4426), p. 19, 1980.
- [11] Y. S. Babu, C. E. Bugg, and W. J. Cook, "Structure of calmodulin refined at 2.2 a resolution.," *Journal of molecular biology* **204**(1), p. 191, 1988.
- [12] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290**(5500), p. 2323, 2000.
- [13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation* **15**(6), p. 13731396, 2003.
- [14] K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proceedings of the twenty-first international conference on Machine learning*, 2004.

A Quick Introduction to Data Compression Through Learning

Francisco Claude

David R. Cheriton School of Computer Science
University of Waterloo, Canada
fclaud@cs.uwaterloo.ca

ABSTRACT

We present an introduction to data compression with an example of a simple technique based on Bayesian Inference, which achieves compression ratio similar to known compression programs in practice. We relate this method to known compression techniques. The main goal is to show data compression from a learning point of view and encourage further research on compression of biological sequences.

1 Introduction

Data compression aims at representing a sequence using as little space as possible. Compression algorithms can be roughly divided into two groups: dictionary based and statistical.

Dictionary based methods can be explained as representing the sequence, based on previously seen substrings. Some examples of such algorithms are LZ77 [1], LZ78 [2] and Re-Pair [3] ¹.

Statistical compressors rely on predicting the next symbol to appear in the sequence, encoding this information and the predictive model in an efficient way. Classical examples are Huffman [4] and arithmetic coding [5]. In some cases, transforming the sequence allows to achieve better compression with simple methods, such an example is the Burrows-Wheeler transform [6, 7].

We first discuss some necessary background in Section 2; then, in Section 3, we show a simple and direct way to use Bayesian inference to compress data. In Section 4 we show experimental results obtained by the methods proposed in Section 3, aiming at giving some empirical feeling to the reader. Then, we discuss the connection of our proposal to existing work (Section 5), and finally we conclude mainly presenting a simple open problem (Section 6).

2 Data Compression

The efficiency of a compression method is usually measured by comparing the length of the resulting sequence with the entropy of the source.

DEFINITION 2.1 (ENTROPY [4]). *The entropy of a sequence S of length n , drawn from an alphabet Σ of size σ , is defined as*

$$H(S) = - \sum_{c \in \Sigma} P(c) \log P(c),$$

where $P(c)$ is the probability of seeing a symbol c in S .

If we consider each symbol independent from each other, the resulting value for the entropy is called zero-order entropy (H_0) and it is a lower bound on a symbol-based coder. This means that we can not compress a sequence S to less than $nH_0(S)$ if we give a single code to each symbol of the alphabet ignoring the context in which it appears. Another definition, very useful in practice, is the k -th order entropy [4], H_k , it considers contexts of length k and is defined as

$$H_k(S) = - \sum_{s \in \Sigma^k} P(s) \sum_{c \in \Sigma} P(c|s) \log P(c|s).$$

In a practical setting this definition is expressed using frequencies, $P(c) = C(c)/n$, where $C(c)$ is the number of times c appears in the sequence. This corresponds to the definition of *empirical entropy* [7].

A simple approach to achieve compression close to the k -th order entropy is to model the symbols distribution for each possible context of length k , assign zero-order codes to each model and then compress the sequence using

¹The last two have a straightforward grammar-representation and are sometimes referred as grammar-based compressors.

them. This method is simple and seems that it could achieve a good compression ratio, but it hides a huge overhead in storing the models. There are σ^k contexts of length k that have to be stored, each of them with their corresponding model.

For the rest of this work we will assume the sequence we want to compress to be generated by a stationary Markov process in which each symbol depends only on the past k symbols seen. So, for our purposes, $H(S)$ and $H_k(S)$ are the same. We make use of the n -gram model used for natural language processing (NLP) [8].

It is clear that there will be a trade off between the complexity of the model, the space required to store it, and the compression ratio achieved. From the pure compression point of view, it is interesting to have a model that adapts itself over the part of the sequence already seen. This allows the encoder and the decoder to adapt their model in a similar way and by doing so encode and decode without storing much of the model (only the initial assumptions or priors). A compression method that works this way is Prediction by partial matching (PPM) [9]. The main idea is to adapt the model based on the context seen, and if that context has not been seen before, the (de)compressor tries to find a shorter context seen before in order to predict the next symbol.

In this work we will rely on Arithmetic compression (AC) [5]. Given a probability distribution of a set of sequences, AC maps the probability of a given sequence to a range in $[0, 1)$, and allows to represent that sequence with a number in that range. In order to avoid the assumption of infinite precision, many authors (see [10] for further details) have showed how to handle the intervals using finite precision for the numbers and improve upon coding efficiency. The most important property of AC for our work is that the encoding process is separated from the probability model we use for the source, and thus it allows us to modify it as we learn from the sequence and not having to deal with the problems presented by other codification methods like Huffman [11] for this scenario. It has been shown that AC achieves compression close to the entropy of the sequence, in particular, we have the following theorem:

THEOREM 2.2 (ARITHMETIC CODING OPTIMALITY [5]). *Let S be a sequence of length n , drawn from an alphabet Σ of size σ . Consider L to be the length of encoding S using arithmetic coding (with the right probability distribution). Then, L satisfies*

$$|S|H(S) < L < |S|(H(S) + 2).$$

There have been other approaches for estimating the probability of a symbol given the history seen in the sequence. In particular, an interesting work that plugs AC with their modeling system was presented by Davies and Moore [12], where they show how to train a Bayesian Network in order to estimate the probability distribution of the next symbol.

One of the best compression methods, considering compression ratio, is PPM [9]. The main idea is to model the source and predict the next symbol based on the symbols seen previously. Usually, we have to define a model for predicting the next symbol and this model is application-dependent [13]. The classical implementation works in a similar way to the two examples shown in this survey, we comment on this in Section 5.

3 Simple Statistical Compressor

In general, the n -gram model for NLP is used over words. In our case, we will apply this over symbol. The same ideas are valid for word-based compressors, which in practice have shown to achieve better compression ratios for natural language [14].

Let $S = s_1 s_2 \dots s_n$ be a sequence of length n , drawn from an alphabet $\Sigma = \{1, 2, \dots, \sigma\}$ of size σ . We will consider contexts of length k . Our goal is to estimate $P(s_j | s_{j-1} s_{j-2} \dots s_{j-k})$. For estimating the probability of $P(s_j | s_{j-1} s_{j-2} \dots s_{j-k})$, we will use a very simple maximum likelihood estimator [8, 15], where we model $P(s_j s_{j-1} \dots s_{j-k})$ as $C(s_j s_{j-1} \dots s_{j-k}) / (N + B)$; N is the number of training instances, B is the number of possible classifications for the training text, and $C(s)$ corresponds to the frequency of s . For estimating the probability $P(s_j | s_{j-1} s_{j-2} \dots s_{j-k})$ we have:

$$P(s_j | s_{j-1} s_{j-2} \dots s_{j-k}) = \frac{P(s_j s_{j-1} s_{j-2} \dots s_{j-k})}{P(s_{j-1} s_{j-2} \dots s_{j-k})} = \frac{C(s_j s_{j-1} s_{j-2} \dots s_{j-k})}{C(s_{j-1} s_{j-2} \dots s_{j-k})}$$

This estimation has problems with zero frequencies, the probability of seeing a new symbol after a context is zero. For fixing this issue, we will use two different methods:

- M0: Laplace Law's results by adding one to the frequency count, obtaining

$$P_{Lap}(s_1 s_2 \dots s_n) = \frac{C(s_1 s_2 \dots s_n) + 1}{N + B}.$$

By using this, we actually obtain the value of the Bayesian estimator that assumes a uniform prior over the symbols in the sequence [8].

- M1: As a second approach, we will consider a special symbol representing an unseen symbol of the alphabet. For the coding purposes, each time we see a symbol whose probability is zero, we emit this special symbol and then we encode the symbol using $\lceil \log \sigma \rceil$ bits. Then we update the frequency counters and the probability distribution for the context.

For both methods, as in general with AC, we add a special symbol to the sequence that represents the end of the string, which is the symbol that tells the decoder when to stop decoding.

4 Experimental Setup

We compare the two models with AC² (M0 and M1) against the compression ratios reported in Pizza&Chili³ for bzip2, gzip, and ppmd. Bzip2 is based on the Burrows-Wheeler Transform [6], gzip is based on the Lempel-Ziv family [1, 2], and ppmd is an implementation of PPM [9]. In this test we consider the file dna.50MB. Bzip2 achieves 25.98%; gzip 27.05%; ppmd achieves 23.84%; finally M0 and M1 achieve 23.84% and 23.82% respectively. This shows that in this case the methods are competitive with general purpose compressors.

The second test considers the file Hemo⁴. We compare the result of compressing this file against compressors designed for the specific task of compressing biological data, using the results provided in [16]. XM [17]⁵ achieves 7.64%, dna2 [18] achieves 9.65%, and Comrad [19] achieves 12.68%. The best lines achieved by M0 and M1 are 17.75% and 17.18%. As a guideline, gzip and bzip achieve 10.81% and 10.77% respectively.

Next, we study the effect of k , the length of contexts, in the compression. Figure 1 (left and middle) shows the compression ratio, as a fraction of the text, for different k s. It can be seen that the function described reaches a minimum and then it starts growing. The problem here is that when we divide the training set into too many bins, which is the case for large contexts, the number of elements hitting the bins are few and the prediction looses. With more data it would be possible to achieve better results for larger contexts, but it is given by a trade off on how much we can collect from the sequence we are compressing. As an example to illustrate this, we consider the DNA file and generate 3 prefixes of it: 10, 30 and 50MB. Each prefix is compressed with different k s and we evaluate the compression ratio achieved. Figure 1 (right) shows the results. As it can be seen, the prediction for larger k s improves in the larger files, and the longer the file, the better the compression tends to be.

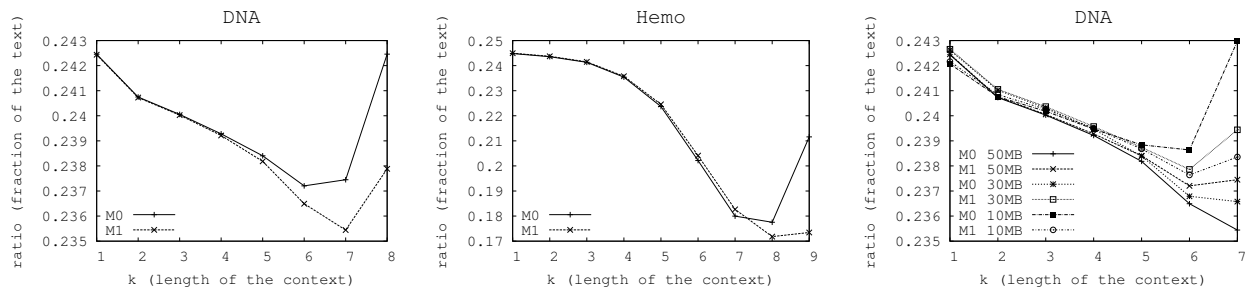


Figure 1. Compression ratio, as fraction of the text, for DNA and Hemo using methods M0 and M1. The leftmost picture shows that the methods achieve better compression as they see more examples.

5 Connection to Other Compression Methods

There is direct connection between the method explored in this work and PPM [9]. The main difference is how PPM handles the non-existing contexts or zero frequencies. The solutions used by PPM is quite similar, if it finds a context that has not been seen before, it tries with a shorter one. It keeps doing this until reaching the point of a

²Using the implementation of [5].

³<http://pizzachili.dcc.uchile.cl/>

⁴<http://ww2.cs.mu.oz.au/~kuruppu/comrad/hemoglobin.fa.gz>

⁵Which uses a much more sophisticated modeler in a similar approach, keeping many modelers and combining them by the use of learning algorithms. In general the compression of biological data is hard because contexts of length k are not a good predictor.

context of length 0 or finding a context that allows to model the next symbol. This seems to be an alternative to the previous model that considers more information for modeling. If a short sequence is very rare, it can be used as a context instead of trying to embed this rare sequence inside a larger context, where it is very likely that we will not find many training examples and thus lose prediction capabilities.

All these methods, the ones presented here, PPM, etc., are just different methods for predicting the next symbol in a sequence plugged with arithmetic coding. This means, we are just trying to find a good alternative for the modeler. In the case of DNA and protein sequences this has been proven to be a very challenging problem.

6 Conclusions and Open Problems

We showed a simple relationship between learning/inference and data compression. This connection is not new, it was previously stated by Cover and Thomas [4], where they relate a good data compressor with a good gambler. In principle the idea is the same, but in our case we limited our selves to Bayesian gamblers.

The option of exploring more general or powerful estimation methods and applications-dependent priors is an exciting path to work further in this topic. Another attractive direction is based on the convergence of the model to the real distribution. A common measure to quantify the closeness between two probability distributions is the Kullback-Leibler distance. In the strict sense it is not a distance, but it is always positive and evaluates to 0 only if the two probability distributions are the same.

DEFINITION 6.1 (KULLBACK-LEIBLER DISTANCE[4]). *The Kullback-Leibler (KL) Distance between two probability distributions $P(x \in X)$ and $Q(x \in X)$ is defined as:*

$$D(P||Q) = \sum_{x \in X} P(x) \frac{P(x)}{Q(x)}$$

It would be really interesting to see how fast can we approach the real probability distribution of the symbols. Given this result, and the following theorem, we could give a bound on the compression achieved by our method.

THEOREM 6.2 (WRONG CODE[4]). *The expected length under $P(x)$ of the code assignment with lengths $l(x) = \left\lceil \frac{1}{Q(x)} \right\rceil$ satisfies:*

$$H(P) + D(P||Q) < E_{Pl}(x) < H(P) + D(P||Q) + 1$$

Finally, another path to explore is to build a transformation based on the estimations made. The main idea is that when processing position $\ell + 1$, we have seen $s_1 s_2 \dots s_\ell$ and we can try to predict $s_{\ell+1}$. If we write down the number of trials required to predict this next symbol. We can recreate the original sequence from this one by running a decoder that does the inverse process. If our predictor is good, we will have a sequence biased on to small numbers, and thus we could aim at compressing it very well. This could lead to similar results as the Burrows-Wheeler transform, where simpler compression methods applied over the transformation achieve competitive results.

It is interesting to notice that the transformation is very similar to the idea used originally by Shannon to estimate the entropy (upper bound) of the English language [20, 21]. The estimation was later improved by Cover and King [22], where they approach the problem as a gambling question, which helps to estimate a probability for the next symbol and not only the order of possible symbols for the next position. An interesting discussion about this can be found in the work presented by Teahan and Cleary [23].

Acknowledgements

The author would like to thank Gonzalo Navarro and Pascal Poupart for their useful comments.

References

- [1] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory* **23**(3), pp. 337–343, 1977.
- [2] J. Ziv and A. Lempel, "Compression of individual sequences via variable length coding," *IEEE Transactions on Information Theory* **24**(5), pp. 530–536, 1978.
- [3] J. Larsson and A. Moffat, "Off-line dictionary-based compression," *Proc. IEEE* **88**(11), pp. 1722–1732, 2000.

- [4] T. Cover and J. Thomas, *Elements of information theory*, John Wiley and Sons, Inc., 1991.
- [5] M. C. E. Bodden and J. Kneis, “Arithmetic coding revealed - a guided tour from theory to praxis,” Tech. Rep. SABLE-TR-2007-5, Sable Research Group, School of Computer Science, McGill University, Montréal, Québec, Canada, May 2007.
- [6] M. Burrows and D. Wheeler, “A block sorting lossless data compression algorithm,” Tech. Rep. Technical Report 124, Digital Equipment Corporation, 1994.
- [7] G. Manzini, “An analysis of the Burrows–Wheeler transform,” *Journal of the ACM* **48**(3), pp. 407–430, 2001.
- [8] C. D. Manning and H. Schtze, *Foundations of Statistical Natural Language Processing*, The MIT Press, June 1999.
- [9] J. Cleary and I. Witten, “Data compression using adaptive coding and partial string matching,” *IEEE Transactions on Communications* **32**, pp. 396–402, Apr 1984.
- [10] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann Publishers, 1999.
- [11] D. Huffman, “A method for the construction of minimum-redundancy codes,” *Proceedings of the I.R.E.* **40**(9), pp. 1090–1101, 1952.
- [12] S. Davies and A. Moore, “Bayesian networks for lossless dataset compression,” in *Proceedings of the Fifth International Conference on Knowledge Discovery in Databases*, pp. 387–391, AAAI Press, 1999.
- [13] D. Mackay, *Information Theory, Inference & Learning Algorithms*, Cambridge University Press, June 2002.
- [14] J. Adiego and P. de la Fuente, “On the use of words as source alphabet symbols in ppm,” in *DCC '06: Proceedings of the Data Compression Conference*, p. 435, IEEE Computer Society, (Washington, DC, USA), 2006.
- [15] F. Peng and D. Schuurmans, *Combining Naive Bayes and n-Gram Language Models for Text Classification*, vol. 2633, January 2003.
- [16] F. Claude, A. Fariña, M. Martínez-Prieto, and G. Navarro, “Compressed q -gram indexing for highly repetitive biological sequences,” in *Proc. 10th IEEE Conference on Bioinformatics and Bioengineering (BIBE)*, 2010. To appear.
- [17] M. Cao, T. Dix, L. Allison, and C. Mears, “A simple statistical algorithm for biological sequence compression,” in *Proc. DCC*, pp. 43–52, 2007.
- [18] G. Manzini and M. Rastreo, “A simple and fast DNA compression algorithm,” *Soft. Pract. Exper.* **34**, pp. 1397–1411, 2004.
- [19] S. Kuruppu, B. Beresford-Smith, T. Conway, and J. Zobel, “Repetition-based compression of large DNA datasets,” in *Proc. 13th International Conference on Computational Molecular Biology (RECOMB)*, 2009. Poster.
- [20] C. E. Shannon, *A Mathematical Theory of Communication*, CSLI Publications, 1948.
- [21] C. E. Shannon, “Prediction and entropy of printed English,” *The Bell System Technical Journal* **30**, pp. 50–64, 1951.
- [22] T. M. Cover and R. C. King, “A convergent gambling estimate of the entropy of English,” *IEEE Transactions on Information Theory* **24**, pp. 413–421, July 1978.
- [23] W. J. Teahan and J. G. Cleary, “The entropy of english using ppm-based models,” in *DCC '96: Proceedings of the Conference on Data Compression*, p. 53, IEEE Computer Society, (Washington, DC, USA), 1996.

High gain lateral amorphous selenium (a-Se) detector for medical imaging

Shiva Abbaszadeh^{a,1}, Kai Wang^a, Nicholas Allec^a, Feng Chen^a, and
Karim S. Karim^a

^a University of Waterloo, 200 University Avenue West, Waterloo, Canada

ABSTRACT

Amorphous selenium (a-Se) is a well known photoconductor and has been used in both indirect and direct conversion x-ray detectors for a variety of medical imaging modalities such as mammography. It goes without saying that interest for having a photodetector with higher gain never ceases. There has been a lot of research on taking advantages of the avalanche multiplication phenomenon inside a vertical a-Se structure to produce high internal gain in the photodetector (e.g. HARP camera for low-light settings). The fast response time and high gain of the a-Se avalanche photodetector makes it a promising candidate to replace photomultiplier tubes (PMT) or silicon avalanche photodiode (APD) based photomultipliers (SiPM) in applications such as positron emission tomography (PET) detectors. Recently, a lateral metal-semiconductor-metal (MSM) a-Se photodetector has been reported as a competitive alternative in terms of ease of fabrication and integration, speed and low dark current. Thus, we believe the lateral structure is also promising for high gain photodetector applications like PET. In this paper, we intend to investigate the effect of increasing electric field on the lateral a-Se structure and compare the results with the modified lucky drift model which presents a good agreement with experimental data on avalanche multiplication in vertical a-Se structures. Our study shows that a gain of ~ 100 can be achieved in a lateral structure under a modest field strength of $40 \text{ V}/\mu\text{m}$. Even though the observed dark current of $\sim 800 \text{ nA}$ is still far beyond the requirement of a practical detector, the achievable high gain allows us to design better detectors for high-end applications such as PET with a unique lateral approach.

Keywords: Amorphous selenium, Photodetector, Avalanche gain

1 INTRODUCTION

Imaging technology always embraces detectors with higher sensitivity and lower noise. For instance, the early detection of breast cancer is crucial for efficient treatment. To produce a high quality image, either the electronic noise should be kept to a minimum or the x-ray photoconductor's conversion gain should be enhanced. For applications like mammography tomosynthesis where the x-ray dose is low, the quantum noise is quite significant [1]. In this case, increasing the photoconductor's conversion gain is the best solution. Vast attention has been given to the avalanche multiplication phenomenon to achieve internal gain inside the photoconductive material. Avalanche photodiodes have been used in many applications such as optical communication. Recently, IBM scientists utilized nanophotonic avalanche photodetectors on a small silicon circuit to replace the electrical signal that is used to communicate through wires between computer chips [2]. Avalanche multiplication in amorphous selenium was reported for the first time by Juska et al. in 1980 [3]. The device structure used by Juska was a simple structure of a-Se sandwiched between two insulating polyethyleneterephthalate layers. The insulating layers were used to avoid any possible charge injection from the electrodes due to the high applied electric field. The insulating layer however prevents the exit of photogenerated carriers to the external circuit. Some years later, using a-Se with proper blocking contacts lead to the commercial deployment of high-gain avalanche rushing photoconductor (HARP) TV camera tubes [4]. The avalanche multiplication converts a faint optical signal to a significant electrical signal which is suitable for low-light level conditions compared to other competing technologies.

Although the existence of avalanche multiplication in a-Se has been known for a relatively long period, the process of avalanche multiplication in amorphous semiconductors is not yet fully understood according to the different motion of electrons due to the inherent disorder potential inside amorphous materials. Among existing

¹ Corresponding author. E-mail: sabbasza@uwaterloo.ca

models intending to describe avalanche multiplication inside amorphous materials, the modified lucky drift model has been shown to provide a good fit to experimental data [5]. The modified lucky drift model considers the effect of carrier scattering due to the disorder potential in addition to carrier scattering with phonons (which is the subject of the lucky drift model for crystalline material). In this model, primary carriers continuously gain and lose energy based on elastic and inelastic scattering in their paths across the electric field. In order to achieve an avalanche state, the carriers should acquire the ionization threshold energy required to excite secondary carriers. The mean free path of carriers in amorphous material is very small (in the range of a few interatomic spaces) [6]. This means the electric field should be sufficiently high to help carriers reach the ionization threshold energy during their transport along their path. Although there is an advantage to increasing the electric field, there is a limitation due to the breakdown of the material caused by the injection of excess carriers. It has been found that the right choice of electrode and blocking materials can reduce this problem in multilayer vertical a-Se structures.

It should be noted that the avalanche multiplication phenomenon is not the only mechanism responsible for achieving photocurrent multiplication. Space charge limited or internal field modification limited processes can lead to photocurrent multiplication. For instance, a gain of up to 70 has been reported in hydrogenated amorphous silicon-based p-i-n junction with an a-SiN:H layer arising from the tunneling of electrons across the band gap through localized states in the a-SiN:H layer [7]. To get photocurrent multiplication by this mechanism, the right choice of layers with proper band gaps and density of localized states is crucial. After illumination, the field redistribution should be in favor of tunneling.

In this work we have fabricated a lateral a-Se photodetector structure and have studied its photoconductive response at 468 nm wavelength illumination. In comparison with different detector technologies, a lateral selenium detector has some advantages which are summarized in Table 1. In particular, advantages include the ease of fabrication and integration, low voltage of operation, speed of operation, and feasibility in a variety of imaging modalities. In addition, a lateral structure appears to easily achieve high gain compared to a vertical avalanche structure in which high field must be applied.

2 Experimental

In this section, we demonstrate experimental results on the investigation of photocurrent characteristics and quantum yield in the lateral a-Se photodetector structure. The cross-sectional diagram of the lateral a-Se photodetector with aluminum electrodes is shown in Fig. 1. This structure was fabricated by conventional microelectronic processes by using two photolithography masks. The details of the fabrication can be found in reference [8]. This design has an electrode spacing of 1 μm , electrode length of 60 μm and electrode width of 2 μm . The thickness of the a-Se layer is 2 μm .

Table 1. Comparison of different technologies.

Detector Factor	Lateral Selenium	Vertical Selenium	P-I-N
Ease of Fabrication	Very Good no blocking contacts	Good	Good
Detection Mode (Spatial Resolution)	Indirect	Direct (best)	Indirect
Voltage of Operation	Low (30-60 V)	High (>2500 V)	Lowest (5-10V)
Ease of Integration	Simple	Simple	Challenging (additional masks)
Speed of Operation	1000 Hz	1-3 Hz	30 – 60 Hz
External Quantum Efficiency	> 1 potentially	< 1	< 1
Avalanche Gain	Possible	Possible	No
Medical Imaging Applications	Fluoroscopy, flat panel CT, Dental Imaging, SPECT & PET	Mammography	Fluoroscopy, flat panel CT, Dental Imaging, SPECT & PET



Figure 1. Cross-sectional diagram of fabricated lateral detector (left) and micrograph of the fabricated detector (right).

The Current-Voltage characteristics of the device with and without illumination were measured (Fig. 2-a). The measurements were carried out using a Keithley S2600 low-noise microprobe station. In order to measure photocurrent, the device was illuminated by a fixed (continuous) incident monochromatic light of wavelength 468 nm. The light intensity was $180 \mu\text{W}/\text{cm}^2$. Fig. 2-b shows the photocurrent at different electric field strengths ranging from $20 \text{ V}/\mu\text{m}$ to $40 \text{ V}/\mu\text{m}$. The photocurrent I_{ph} is given by the subtraction of dark current from the measured current under illumination. The quantum yield of photogeneration was calculated using the measured photocurrent by the following expression:

$$\eta = \frac{I_{\text{ph}}/e}{I/h\nu} \quad (1)$$

where e is the elementary charge, and $h\nu$ is the energy of incident photon.

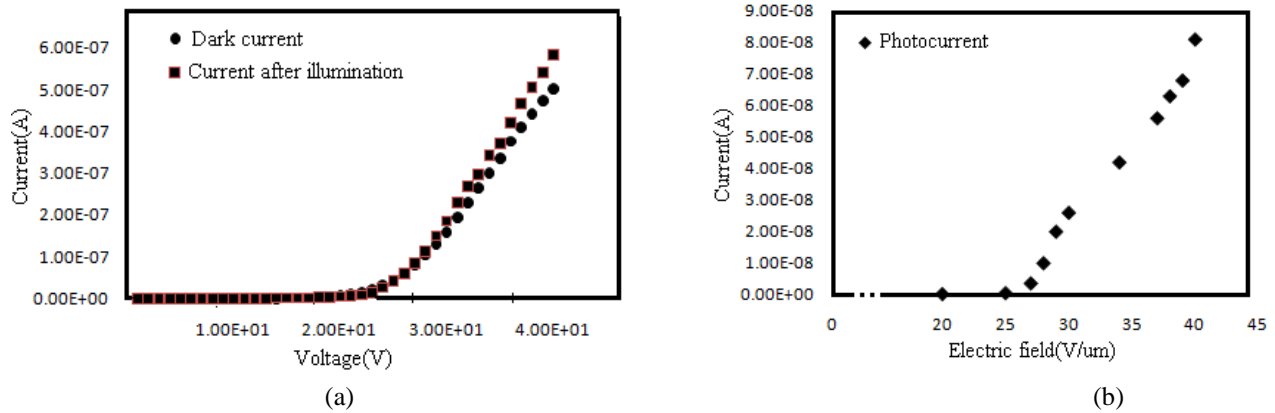


Figure 2. a. Signal current and dark current versus voltage, b. photocurrent-field characteristic.

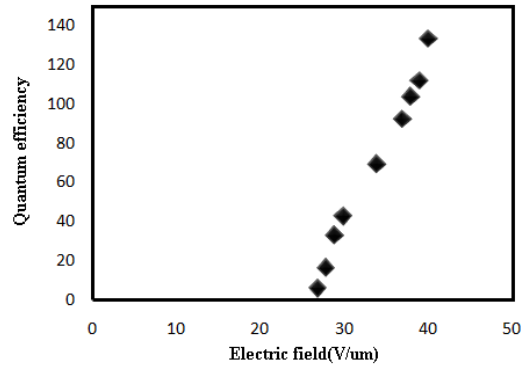


Figure 3. The dependence of quantum yield on applied electric field.

Fig. 3 shows the calculated quantum yield as function of electric field. The quantum yield for an electric field of $25 \text{ V}/\mu\text{m}$ was close to unity. We calculated the gain (G) by the following relation:

$$G = \eta(E)/\eta(25\text{V}/\mu\text{m}) \quad (2)$$

where $\eta(25\text{V}/\mu\text{m})$ is approximately equal to unity. Therefore, the gain and quantum yield for electric field above $25\text{V}/\mu\text{m}$ are the same. The lucky drift model predicts an avalanche gain for electric field larger than $70\text{-}80 \text{ V}/\mu\text{m}$ in

vertical structure. This result suggests that the avalanche multiplication is not the responsible mechanism for creating gain in this structure. One of the possible mechanisms responsible for the excess photocurrent and resulting gain might be due to the injection of more holes from the electrode [9]. The hole mobility ($\sim 0.12\text{cm}^2/\text{Vs}$) in a-Se is generally two orders of magnitude higher than the electron mobility ($\sim 0.003\text{cm}^2/\text{Vs}$) [10]. After generation of electron-hole pairs due to incident light, while the faster holes are collected by the cathode, the electrons are still proceeding towards the anode. The process creates an absence of holes and hence a net negative charge in the region that is compensated for by injection of holes from the anode into this region. This process leads to the generation of more holes upon absorption of a single photon. We are working to better understand the existence of gain in this type of device structure, including exploring other possible mechanisms responsible for the observed gain.

3 Conclusion

A photocurrent multiplication phenomenon is observed in a lateral a-Se based photodetector. A quantum yield of 136 is obtained in the presence of an applied electric field of $40\text{ V}/\mu\text{m}$ which eliminates the need of applying a very high bias potential across the photodetector. The results suggest that the gain is not due to avalanche multiplication, which is the responsible mechanism for carrier multiplication in vertical structures. Currently, the main drawback of the device is high dark current. This point needs to be improved before the device can be used for PET detector.

ACKNOWLEDGMENTS

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- [1] L. Romualdo, M. Vieira, and H. Schiabel, "Mammography Images Restoration by Quantum Noise Reduction and Inverse MTF Filtering", *Computer Graphics and Image Processing, Symposium*, pp. 180-185, 2009.
- [2] S. Assefa, F. Xia, and Y. A. Vlasov, "Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects", *Natur.* 464, pp. 80-84, 2010.
- [3] G. Juska, and K. Arlauskas, "Impact Ionization and Mobilities of Charge Carrier at High Electric Fields in Amorphous Selenium", *Phys. Stat. Sol. (a)* 59, pp. 389-393, 1980.
- [4] K. Tanioka, J. Yamazaki, K. Shidara, K. Taketoshi, and T. Kawamura, "Avalanche-mode Amorphous Selenium Photoconductive Target for Camera Tube", *Proc. Photo-electronic image device*.100, pp. 379-385, 1988.
- [5] O. Rubel, S. D. Baranovskii, I. P. Zvyagin, P. Thomas, and S. O. Kasap, "Lucky-drift for avalanche multiplication in amorphous semiconductor", *Phys. Stat Sol.(c)* 1, pp. 1186-1193, 2004.
- [6] S. O. Kasap, J. A. Rowlands, S. D. Baranovskii, K. Tanioka, "Lucky drift impact ionization in amorphous semiconductors", *J. Appl. Phys.* 96, pp. 2037-2048, 2004.
- [7] M. Yoshimi, T. Ishiko, K. Hattori, H. Okamoto, and Y. Hamakawa, "Photocurrent multiplication in hydrogenated amorphous silicon-based p-i-n junction with an a-SiN:H layer", *J. Appl. Phys.* 72, pp. 3186-3193, 1992.
- [8] K. Wang, F. Chen, G. Belev, S. O. Kasap, K. S. Karim, "Design and modeling of lateral a-Se MSM photoconductor as indirect conversion X-ray imager", *Proc. SPIE.* 7449, 74491W, 2009.
- [9] R. P. Khare, *Fiber Optics and Optoelectronics*, Oxford University Press, 2004.
- [10] M. Yunus, "Monte Carlo Modeling of the sensitivity of X-ray photoconductors", M.Sc. Thesis, University of Saskatchewan, Saskatoon, Canada, 2005.

An *in silico* mathematical model of the initiation of DNA replication

Rohan D. Gidvani^{a1}, Brendan J. McConkey^a, Bernard P. Duncker^a and Brian P. Ingalls^{a,b}

^aDepartment of Biology, University of Waterloo, 200 University Ave. W, Waterloo, Canada;

^bDepartment of Applied Mathematics, University of Waterloo, 200 University Ave. W,
Waterloo, Canada

ABSTRACT

Proper eukaryotic cell proliferation depends upon DNA replication, a closely regulated process mediated by the actions of a multitude of factors. The initiation of replication is regulated by the heterohexameric Origin Recognition Complex (ORC). At origins of replication, ORC recruits and/or associates with protein factors such as Cdt1, Cdc6, the MCM2-7 complex, Cdc45 and the Dbf4-Cdc7 kinase. The mechanisms controlling these associations are well documented, allowing the development of a mathematical model that allows us to explore the network's behaviour. Using budding yeast as a model organism, we have developed an ordinary differential equation (ODE)-based model of the protein-protein interaction network regulating replication initiation. Precise quantification of protein factors at various timepoints is critical to calibration of the model parameters. To this end, we have made use of genetic manipulations and quantitative protein expression analysis. Using chromatin extracts from synchronized cell cultures, we were able to monitor the fluctuation of a number of the aforementioned proteins. This information was used to infer qualities of the protein network and to calibrate a predictive mathematical model of the process of DNA replication initiation, which can be integrated into existing models of the entire budding yeast cell cycle.

Keywords: DNA replication; origin recognition complex; pre-replicative complex; mathematical modeling, systems biology, cell cycle

1 INTRODUCTION

The machinery of the eukaryotic cell cycle has been extensively dissected and described from simple to complex organisms. Cell proliferation hinges on the ability to replicate the genome with high fidelity, segregate the chromosomes equally and finally divide the cell, resulting in two genetically identical copies. Equally important are the monitoring modules that oversee these pathways and that intervene under unfavourable conditions, such as DNA damage, and trigger the ensuing repair mechanisms. These steps have been extensively characterized and the cycle organized into an approximate pathway of sequential events.

In eukaryotes, a functionally-conserved heterohexameric protein complex – ORC (Origin Recognition Complex) acts as a selector for origins of DNA replication. ORC then serves as a scaffold for the association of a number of additional replication factors, which collectively form the pre-replicative complex (Pre-RC). The protein encoded by the *CDC6* gene is also essential and is required for initiation via its role in loading the heterohexameric MCM (minichromosome maintenance) complex onto origin DNA. The six subunits, Mcm2-7, when formed into an active complex, collectively act as the replicative helicase. Formed in the cytoplasm, the MCM complex is co-transported to the nucleus with Cdt1 and is recruited to the ORC- and Cdc6-bound DNA (reviewed in [1]). This is promoted by the direct interaction between Orc6 and Cdt1 [2]. These steps culminate in the loading of the MCM rings onto DNA, whereupon they unwind the double helix bidirectionally and provide

¹ Corresponding author. Department of Biology, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1. Tel: +1 519 888 4567 x36551; Email: rdgidvan@uwaterloo.ca

access for the DNA polymerases. This tight loading is dependent on a stepwise ATP-hydrolysis dependent mechanism involving Cdc6 and ORC [3].

Pre-RCs are set up at about 300 of the approximately 500 consensus binding sequences, to define potential origins. Firing of a particular origin is dependent upon the association of another group of proteins resulting in the formation of the Pre-IC (Pre-Initiation Complex). Cdc45, the GINS complex, Sld2, Sld3 and Dpb11 must be recruited to a licensed origin, with Cdc45 being a limiting factor. Crucial to this step is the phosphorylation of a number of these proteins by the cyclin-dependent kinase, Cdk1. This is the common name generally referring to the combination of the protein kinase Cdc28 with a cyclin (in the case of DNA replication, cyclin-B5, the gene product of *CLB5*) thus providing a controlling input required for passage through the cell cycle. Finally a second kinase is required primarily to phosphorylate various MCM subunits, activating the ring and triggering initiation. This protein, Cdc7 is also controlled by a limiting regulator: Dbf4. Together they form a kinase complex commonly referred to as Dbf4-dependent kinase (DDK). Dbf4 expression is constitutive, but its degradation is controlled by the anaphase-promoting complex or APC. These mechanisms are reviewed in [1].

In order to maintain genomic stability and prevent over- or under-replication of the genome, the cell has evolved mechanisms to ensure that replication occurs exactly once per cell cycle. This is paramount to avoiding loss of genome integrity and/or cell viability [4]. The inhibitory effects of CDKs (promoted by the abundance of Clb5) on Pre-RC components in *S. cerevisiae* are well documented, and ultimately manifest as deactivation, degradation or nuclear export of these factors. Thus the cell cycle exhibits a dual-state behaviour. When Clb5 levels and consequently CDK activity is high, Pre-RCs cannot be established. Once an origin fires in a DDK- and CDK-dependent manner, a new Pre-RC cannot be established until the next cell cycle due to the inhibitory effects of various CDKs, whose activities peak at S phase and remain high until the end of mitosis.

The complex yet elegant network of cell cycle proteins is sophisticated enough to warrant an attempt at modeling, both because many of its key steps are known and because of the inherent difficulty in intuitively determining individual protein behaviours and interactions under varying biological circumstances. Not only does our model seek to elucidate the fundamentals of DNA replication initiation in yeast, but it also strives to attain predictive power. Given that many of the replication factors as well as the processes that oversee their functions are highly conserved from budding yeast to humans, the model has the potential to be extrapolated to attend to biomedical questions. As defects in the cell cycle and particularly DNA can give rise to human cancers, a predictive model of the cell cycle is invaluable in designing targeted cancer therapies and in determining potential side effects. An added level of rigor is provided by the relationship between our model and that of the whole yeast cell cycle [5] using levels of key cell cycle determinants as specified by the latter.

2 METHODS AND RESULTS

2.1 Building a Kinetic Model

In deconstructing a biological system, a network must be created to represent the key proteins involved and the salient interactions. Passage through replication can be thought to begin at the point of ORC binding DNA. This represents our first “species”, which is a relevant unit in the pathway whose abundance changes over time. As described, multiple proteins bind sequentially and exert their well-defined function. The network diagram is shown in Figure 1. This allows us to intuitively separate and examine the various steps. By measuring the abundance of a given protein at a particular point in the cycle, we can ascertain the levels of the various species of which it is a component. In addition, by defining the enzymatic reactions linking the multiple species in terms of parameters and protein concentrations, we can monitor flux through the pathway.

2.2 *In silico* Modeling

Our network is justified by information regarding the qualitative nature of the individual reactions. Our goal was to convert this knowledge into a consensus picture of the molecular reactions as defined by a set of nonlinear, ordinary differential equations (ODE). A simulated annealing algorithm was used to calibrate the models behaviour by minimizing the least-squares-error in comparison with experimental data.

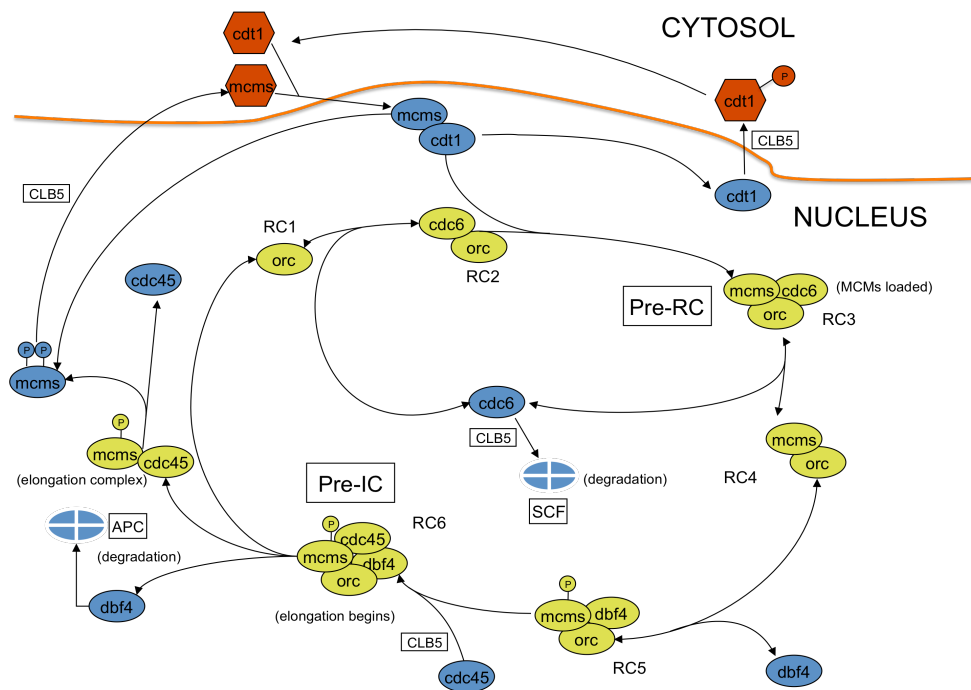


Figure 1. A consensus picture of the network describing the initiation of DNA replication. Reactions are modeled with Mass Action kinetics. The model consists of 11 independent state variables and involves 19 parameters. The levels of APC and CLB5 (representative of CDK activity) are taken as time-varying inputs, with values according to the cell cycle model of Chen et al. [5]. The ODEs are arrived at using the reaction rates shown in Table 1.

2.3 Accumulation of *in vivo* data

In order to determine concentrations of individual proteins implicated in our consensus model we implemented the following methods:

Logarithmically growing asynchronous yeast cultures of a number of strains were arrested in late G1 phase using α -factor. For Cdc6 and Cdc45, myc-epitope tags were incorporated into the open reading frames of these genes and the resultant fusion protein abundances were assayed in separate trials, but by the same method. Cells were released from the G1 block synchronously into the cell cycle. This was confirmed by fluorescence activated cell sorting (FACS). Samples were taken at specific time intervals and were processed by chromatin fractionation to separate proteins bound to chromatin from those that were not. Samples were analyzed by Western blotting, using antibodies directed to the myc-tag or to the protein itself in the case of Mcm2, Dbf4 and Orc2. The amount of a particular protein bound to chromatin (making the assumption that this represents Pre-RC/Pre-IC inclusion) was measured as was the amount unbound. Information about Cdt1 behaviour was obtained from published data [6]. For each protein, at least three trials were performed.

Protein concentrations were determined first by densitometry of Western blots followed by normalization to the number of molecules/cell determined by GFP-tagging experiments described in [8].

2.4 Model Fitting

We used Western blot analysis to measure the abundance of Cdc6, Cdc45, Mcm2 and Dbf4 at eight time points along the cell cycle. This data was compared with the model output and a simulated annealing algorithm was run to minimize the associated least square error associated (normalized by the experimental variance). Figure 2 illustrates the fluctuation of species that comprise the network over time. Figure 3 shows representative examples of the resulting best-fit behaviour.

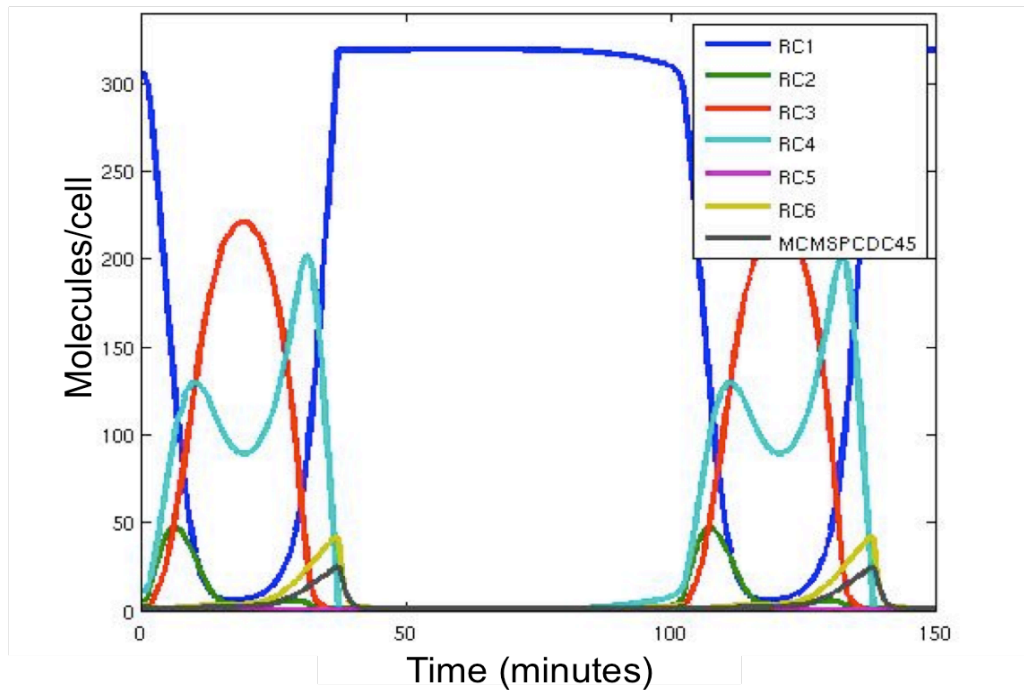


Figure 2. Abundance of network species over time in a 100 minute cell cycle.

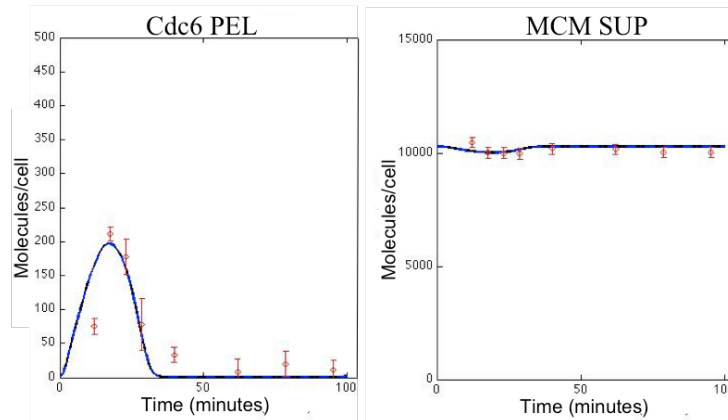


Figure 3. Model simulations of given protein abundances in the indicated cellular fraction (solid line) fit to data points (dots) using a best-fit parameter set. Error bars represent trial-to-trial variation in the data set. The same units used in Figure 1 apply for the respective axes. PEL (pellet) denotes chromatin-bound protein whereas SUP (supernatant) refers to non-chromatin bound protein.

Table 1. Reaction rates. State variables corresponding to the species in Figure 1 are indicated in uppercase.

Description	Rate reaction
Expression of CDC6:	k_1
Degradation of CDC6:	$k_2.CLB5.CDC6$
Expression of MCMS:	k_3
Degradation of MCMS:	$k_4.MCM\ SUP$
Expression of DBF4:	k_5
Degradation of DBF4:	$k_6.DBF4.APC$
Expression of CDC45:	k_7
Degradation of CDC45:	$k_8.CDC45$
<i>Formation of the Pre-RC</i>	
Association of ORC and CDC6:	$k_9.RC1.CDC6$

Association and nuclear import of MCMS and CDT1:	k_10.MCM SUP.CDT1 SUP
Loading of MCMS by CDT1:	k_11.RC2.MCM.CDT1
Nuclear export of CDT1:	k_12.CLB5.CDT1
Dissociation/re-association of nuclear MCM-CDT1 complex:	k_13.MCM.CDT1 - k_14.CDT1.MCM PEL
<i>Formation of the initiation complex</i>	
Dissociation/re-association of CDC6 from the pre-RC:	k_14.RC3 - k_15.CDC6.RC4
Association/dissociation of DBF4 and the pre-RC:	k_16.RC4.DBF4 - k_17.RC5
Association of CDC45 and the pre-RC:	k_18.RC5.CDC45.CLB5

3 DISCUSSION

Our consensus model has produced a high level of matching with the experimental data, suggesting that it represents a good estimation of the network behaviour. In addition to validation against wild-type protein levels, we also compared the model's behaviour to reports of system behaviour under various perturbations (gene knockdowns, shutoffs, over-expressions). Based on the observation that Dbf4 is degraded by the APC and that this process is one of the redundant mechanisms that prevents re-firing of origins, we simulated Dbf4 as being refractory to degradation and observed an increase in the flux through the fork initiation portion of the cycle. Tanaka and Diffley [6] observed co-transport of Cdt1 and MCMs into the nucleus and we mimicked their experiment which abolished transport of either by preventing the other from entering the nucleus. The abundance of the ORC-Cdc6-MCM complex was reduced drastically in simulations. Finally, given that a major driving force of the network is the initial Cdc6-DNA-ORC binding event, increasing its abundance would be expected to have a profound effect on origin firing. Notwithstanding, re-replication is prevented by degradation of Cdc6 by CDKs. There is a reciprocal relationship in that Cdc6 inhibits CDKs themselves (reviewed in [7]). By increasing the amount of Cdc6 by lowering CDK levels (which additionally maintain a high level of MCMs in the nucleus) in our model we were able to produce a cycle in which origins were firing continuously (i.e. re-replication) indicated by the presence of Pre-RCs and replication forks simultaneously. In addition to results from the literature, we are also undertaking our own *in vivo* perturbation experiments to provide further validation of the model. This will allow greater predictive power in using the model to investigate wild-type and disrupted cell-cycle behaviour.

4 REFERENCES

1. Bell, S. P. & Dutta, A. DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* **71**, 333-374 (2002).
2. Chen, S., de Vries, M. A. & Bell, S. P. Orc6 is required for dynamic recruitment of Cdt1 during repeated Mcm2-7 loading. *Genes Dev.* **21**, 2897-2907 (2007).
3. Randell, J. C., Bowers, J. L., Rodriguez, H. K. & Bell, S. P. Sequential ATP hydrolysis by Cdc6 and ORC directs loading of the Mcm2-7 helicase. *Mol. Cell* **21**, 29-39 (2006).
4. Green, B. M. & Li, J. J. Loss of rereplication control in *Saccharomyces cerevisiae* results in extensive DNA damage. *Mol. Biol. Cell* **16**, 421-432 (2005).
5. Chen, K. C. *et al.* Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell* **11**, 369-391 (2000).
6. Tanaka, S. & Diffley, J. F. Interdependent nuclear accumulation of budding yeast Cdt1 and Mcm2-7 during G1 phase. *Nat. Cell Biol.* **4**, 198-207 (2002).
7. Honey, S. & Futcher, B. Roles of the CDK phosphorylation sites of yeast Cdc6 in chromatin binding and rereplication. *Mol. Biol. Cell* **18**, 1324-1336 (2007).
8. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691 (2003).

On sample allocation in differential in-gel electrophoresis: two's a couple, three's a crowd?

Owen Z. Woody^{a1}, Lorna Deeth^b, Catherine Walden^a, Thomas D. Singer^a and
Brendan J. McConkey^a

^aUniversity of Waterloo, 200 University Avenue West, Waterloo, Canada

^bUniversity of Guelph, 50 Stone Road East, Guelph, Canada

ABSTRACT

Historically, differential in-gel electrophoresis (DIGE) experiments have employed a design that incorporates three samples on each gel: two experimental samples and one aliquot of a common reference sample. To counteract the substantial gel-to-gel variability inherent to the gel technology, each experimental measurement was typically expressed as a ratio relative to the common reference, canceling out many gel-specific effects. However, it was recently shown that this approach introduces a bias. Specifically, dividing each of the experimental sample measurements by their co-run reference measurement causes the values to become correlated, violating an assumption of the commonly employed Student's independent samples t-test. As a remedy, it has been suggested that all future experiments incorporate only two samples per gel (experiment & reference), doubling the number of gels required to analyze the same number of experimental samples. Here, we investigate using an alternative statistical test capable of handling correlated data – the dependent samples paired t-test. We show that this test can be used to salvage results from old three-dye experiments. Furthermore, we show that the paired t-test permits analysis without a reference sample in experiments where the number of treatments is small. Potential alternative uses for reference channel measurements are also investigated.

Keywords: Differential in-gel electrophoresis, Statistical analysis of high-throughput data, Experimental design, Reference sample, Simulation

ACKNOWLEDGMENTS

This work was funded by an NSERC CGS-D scholarship awarded to O. Woody.

¹ Corresponding author. E-mail: owoody@uwaterloo.ca, Telephone: +1(519)885-1211 ext. 36686.

Evaluation of New Parameters for Assessment of Stroke Impairment

Kathrin Tyryshkin^a, Janice I. Glasgow^a and Stephen H. Scott^b

^a School of Computing, Queen's University, Kingston, ON, Canada;

^b Department of Anatomy and Cell Biology, Queens University, Kingston, ON, Canada;

ABSTRACT

This paper presents new parameters for the evaluation of stroke impairment using data collected with KINARM robot (Kinesiographical Instrument for Normal and Altered Reaching Movements). The new parameters evaluated in this study were cross-correlation, low frequency, and high frequency. The data were collected from control (people with no neurological disorders) and stroke subjects performing a center outreach task. In this task the subjects were instructed to move the examined arm quickly and accurately from the central target position to a randomly illuminated target, and to maintain the hand at this target for the remainder of the trial. The collected data for each of eight individual reaching movements to eight different targets can be viewed as a time series. For each subject a cross-correlation between the reaching movement to each of eight targets and a straight line fitted between these targets was computed. In addition, high and low frequencies were calculated from the time series data using a Fourier transform. The results showed that the new parameters identified the same or a higher percentage of stroke participants as abnormal, compared to previously reported parameters [1], especially in the experiments performed with the non-affected arm. Therefore, the new parameters can facilitate the detection of abnormalities in the movements of stroke patients and may be used as features for the classification of stroke patients.

Keywords: stroke rehabilitation, cross-correlation, time series analysis, classification

1 INTRODUCTION

A stroke is an acute injury of the brain that can affect many body functions, often causing motor, speech, memory, vision and other sensory impairments. Rehabilitation is an important part of stroke recovery and the key to successful rehabilitation is an accurate assessment of stroke-caused impairment [2, 3]. Current clinical assessments generally involve physical assessment and visual observation by physicians. Therefore, assessment results are inherently subjective and potentially inconsistent among physicians. Moreover, current assessment tools are not adequate to reliably discriminate between different levels of performance. Thus, in practice the majority of stroke patients follow the same general rehabilitation program, which may not necessarily be optimal for each individual case.

Robotics technology can objectively monitor a subject's performance in a given task and even modify the physics of limb motion. The technology can be used in building computational systems that analyze, visualize and aid in the interpretation of sensory-motor dysfunction in stroke patients.

KINARM (Kinesiographical Instrument for Normal and Altered Reaching Movements) is a robotic device developed to study fundamental issues in motor control and learning in upper limbs of primates [4] and is currently used in clinical research for assessing sensor-motor function of stroke patients prior to and during recovery. KINARM allows for the collection of quantitative measurements of upper limb movements of a subject performing a particular task. The collected kinematic and kinetic data are then assessed and stored in a database. The stored data includes measurements such as hand trajectory, elbow position and shoulder angles. From these measurements various additional quantities are derived, among which the initial direction error (IDE)¹ was identified as the best parameter to identify the largest number of stroke participants as abnormal [1]. In this paper, time series analysis was applied on data collected with KINARM. The new parameters evaluated in this study

¹ IDE (in degrees) is an angular deviation between (a) a straight line from the hand position at movement onset to the peripheral target and (b) a vector from the hand position at movement onset to the hand position after the initial phase of movement [1].

were cross-correlation, low frequency, and high frequency. The purpose was to investigate whether the new derived quantities enable better separation of stroke from control subjects. High cross-correlation would indicate that the reaching movements were very close to a straight line. Similarly, a reaching movement with very quick changes and variations would result in a high level of high frequency activity.

2 METHODS

2.1 Data Collection

The data were collected from 39 stroke subjects (22 of which had left arm impairment and 17 had right arm impairment) and 45 age-matched control subjects (people with no neurological disorders). Each subject underwent a typical stroke assessment and one “KINARM session”, where several tasks were performed for each arm. The task of interest for this paper is the center outreach task.

In the center outreach task the subject starts each trial by maintaining the hand at a central target (Fig. 1a). After 1–1.5 seconds, one of eight peripheral targets is illuminated and, as earlier instructed, the subject moves the examined arm as quickly and accurately as possible to the illuminated target. The subject has three seconds to complete the movement and, when at the peripheral target, must maintain the hand at the target for the remainder of the trial. Eight trials were recorded for each target in a pseudo-random blocked design ($n=64$).

While the subject performed this task with each arm, the robot collected quantitative measurements of the movements of each upper limb. The collected data for each of eight individual reaching movements to eight different targets can be viewed as a time series where the X and Y coordinates of the subject’s hand are displayed as a function of the duration of the center outreach task (Fig. 1b).

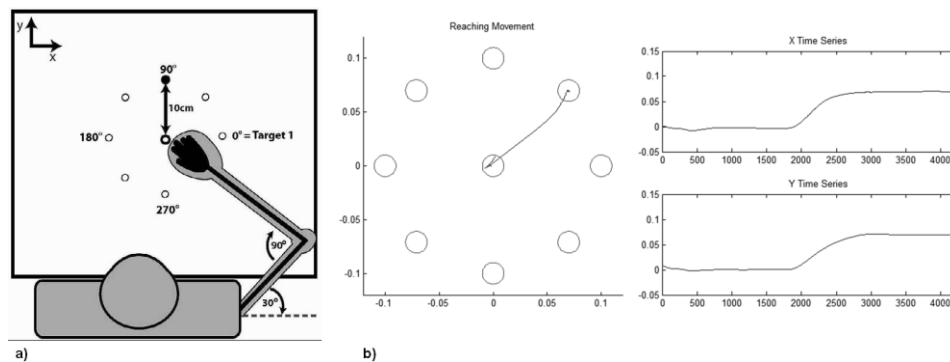


Figure 1. a) The center outreach task. In this task the subject is asked to move one hand from the center position to one of eight targets at which a light is turned on. b) The collected data for each of eight individual reaching movements to eight different targets was interpreted as a time series.

2.3 Data Preprocessing

The time series data were transformed to a reference coordinate system and all eight reaching movements for all eight targets were rotated to match the reaching movement from the central target (T0) to the second target (T2). For the cross-correlation, a set of points representing a straight line from T0 to T2 was created. For the calculation of the high and low frequencies, each reaching movement data set was padded with zeros to ensure that there are power of two data points.

2.4 Parameters Extraction

The zero-lag cross-correlation between a straight line and each reaching movement was computed for all eight trials. The high and low frequencies were derived by summarizing the high and low components of the power spectrum computed for each reaching movement. Similarly to cross-correlation, the high and low frequencies of each reaching movement were computed for all eight trials.

The resulting data contained 80 values for each X and Y coordinate for each subject. The X and Y coordinates were combined using the geometric mean. From the resulting values, the following parameters were computed: (1) mean of eight reaching movements of the mean eight trials (meanOfMeans), (2) median of eight reaching movements of the mean eight trials (medianOfMeans), (3) mean of eight reaching movements of the median of eight trials (meanOfMedians), (4) median of eight reaching movements of the median of eight trials

(medianOfMedians), (5) maximum of eight reaching movements of the maximum of eight trials (maxOfMax), (6) minimum of eight reaching movements of the minimum of eight trials (minOfMin), (7) mean of eight reaching movements of the minimum of eight trials (meanOfMin), (8) median of eight reaching movements of the minimum of eight trials (medianOfMin), (9) mean of eight reaching movements of the maximum of eight trials (meanOfMax), (10) median of eight reaching movements of the maximum of eight trials (medianOfMax), and (11) maximum standard deviation of eight reaching movements for all eight trials (maxOfstd).

3 RESULTS

Fig. 2 compares typical reaching trajectories of the right (dominant) hand of a control subject (on the left) with reaching trajectories of a non-stroke-affected hand of a stroke subject (on the right). The figure clearly shows that some stroke patients do experience difficulties in motor performance not only on their stroke-affected side, but also on their non-stroke-affected side.

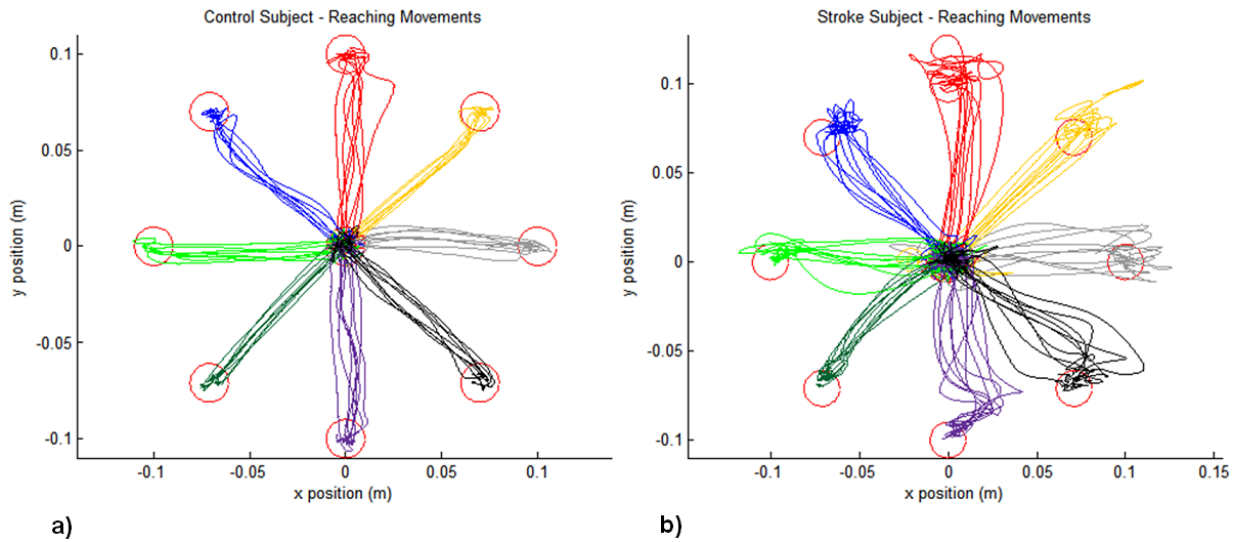


Figure 2. Reaching trajectories: a) control subject (right, dominant limb). b) stroke subject (right, dominant, non-stroke-affected limb).

To compare the performance of stroke subjects with respect to controls using the newly proposed features, the 5% – 95% inter-quartile range of the control subjects was computed for all parameters. Tables 1 and 2 show the percentage of stroke subjects who differed from normal behavior and fell outside the 5% – 95% inter-quartile range of the control group for the center outreach task performed with right and left hand respectively. The results are divided into right affected (RA) and left affected (LA) categories. For comparison, results of the IDE parameter are also shown. Among the new parameters, values that have a higher percentage than the IDE parameter are highlighted in red.

The best results were combined into one feature vector of length 11 that included: meanOfMedians, meanOfMin and maxOfstd of cross-correlation results; and minOfMin, maxOfMax, medianOfMax and maxOfstd of high and low frequency results. A support vector machine classifier was used to separate the resulting data into stroke and control groups. Table 3 shows the ten-fold cross validation classification results with close to 80% accuracy for both the right and left hand experiments.

Finally, the classification results were also compared to the arm and hand Chedoke McMaster clinical scores obtained for both hands on the same day as the center outreach task was performed. The Chedoke-McMaster Stroke Assessment is a questionnaire based, widely used clinical assessment of stroke impairment that maps the physical impairments and disabilities that impact the daily activities of individuals with stroke on a seven-point scale [5]. The results show that a large number of stroke subjects with a perfect Chedoke McMaster score have motor deficits and were successfully classified as ‘stroke’ using the proposed parameters.

Table 1. Percentage of stroke subjects who differed from normal behavior and fell outside the 5% – 95% inter-quartile range of the control group for the center outreach task performed with *right hand*. Results are divided into right affected (RA) and left affected (LA) categories.

		High Frequency		Low Frequency		Cross-correlation	
		RA%	LA%	RA%	LA%	RA%	LA%
1	meanOfMeans	47.06	27.27	41.18	22.73	41.18	36.36
2	medianOfMeans	29.41	27.27	35.29	27.27	41.18	31.82
3	meanOfMedians	23.53	13.64	17.65	22.73	47.06	31.82
4	medianOfMedians	29.41	27.27	29.41	31.82	47.06	36.36
5	maxOfMax	29.41	45.45	35.29	59.09	52.94	45.45
6	minOfMin	64.71	31.82	58.82	45.45	47.06	36.36
7	meanOfMin	35.29	40.91	35.29	27.27	52.94	50
8	medianOfMin	5.88	31.82	17.65	40.91	35.29	31.82
9	meanOfMax	47.06	50.00	35.29	50.00	52.94	40.91
10	medianOfMax	35.29	36.36	41.18	50.00	29.41	31.82
11	maxOfstd	47.06	45.45	47.06	40.91	29.41	22.73
	mean IDE	64.71	36.36				

Table 2. Percentage of stroke subjects who differed from normal behavior and fell outside the 5% – 95% inter-quartile range of the control group for the center outreach task performed with *left hand*. Results are divided into right affected (RA) and left affected (LA) categories.

		High Frequency		Low Frequency		Cross-correlation	
		RA%	LA%	RA%	LA%	RA%	LA%
1	meanOfMeans	11.76	31.82	17.65	40.91	41.18	68.18
2	medianOfMeans	23.53	13.64	23.53	36.36	11.77	50
3	meanOfMedians	17.65	4.55	17.65	31.82	41.18	72.73
4	medianOfMedians	11.76	4.55	29.41	18.18	17.65	63.64
5	maxOfMax	35.29	95.45	35.29	90.91	29.41	68.18
6	minOfMin	17.65	18.18	29.41	45.45	17.65	59.09
7	meanOfMin	11.76	27.27	23.53	31.82	41.18	81.82
8	medianOfMin	11.76	50.00	17.65	45.45	29.41	81.82
9	meanOfMax	41.18	72.73	29.41	68.18	41.18	68.18
10	medianOfMax	41.18	77.27	35.29	68.18	0	36.36
11	maxOfstd	29.41	59.09	35.29	63.64	11.77	40.91
	mean IDE	23.53	72.73				

Table 3. Classification results using a support vector machine classifier and comparison of the results to clinical scores.

Classification			Proposed method vs. Chedoke-McMaster			
	RH (%)	LH (%)	Stroke subjects with perfect arm and hand Chedoke-McMaster score that were <u>correctly</u> classified as “Stroke”			
Correct Rate	80	81	affected limb (n = 13)		non-affected limb (n=33)	
Sensitivity	82	85	RH	LH	RH	LH
Specificity	80	74	10	8	25	22

4 DISCUSSION AND CONCLUSIONS

The results showed that the new parameters identified the same or a higher percentage of stroke participants as abnormal, compared to the IDE parameter, especially in the experiments performed with the non-affected arm. The main advantage of the new parameters over the existing parameters is that they capture the entire nature of the movement as opposed to the IDE parameter which only characterizes the initial phase of the movement of a subject. It only captures the period from movement onset to the first minimum hand speed, which is the first local minimum after the first maximum hand speed. The first minimum hand speed is generally reached within the first

second of the total movement duration. The cross-correlation is a measure of how straight the movement was. A higher value for the cross-correlation indicates that the movement was more accurately directed towards the illuminated target. High standard deviation between trials indicates inconsistent movements. Reaching movement with very quick changes and variations would result in a high level of high frequency activity.

The percentage of left-affected performing with their non-affected and affected limbs is higher than the right-affected subjects performing with their non-affected and affected limbs respectively. This indicates that left-affected subjects tend to show greater deficits in performance with both their affected limbs as compared with right-affected subjects. This pattern was previously detected through other parameters [1].

Some of the new parameters were able to detect more abnormalities in left-affected stroke subjects performing with their non-affected limb (see table 1). A support vector machine was able to separate stroke and control subjects with high accuracy. In addition, classification using the new parameters has a better ability to detect movement abnormalities than the Chedoke-McMaster assessment. Many stroke patients with perfect Chedoke-McMaster scores nonetheless showed abnormal reaching movements and were thus classified as stroke patient. Without the newly proposed parameters, these patients most probably would not receive any rehabilitation for their non-stroke-affected arm and hand. Regardless of the progress achieved with this study, it is important to note that more work needs to be done to more successfully detect motor deficits in stroke patients.

ACKNOWLEDGMENTS

We would like to thank Helen Bretzke and Kim Moore for the assistance with the database and the KINARM data. In addition, we would like to thank Canadian Institute of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting this project.

References

- [1] A. M. Coderre, A.A. Zeid, S.P. Dukelow, M.J. Demmers, K.D. Moore, M.J. Demers, H. Bretzke, T.M. Herter, J.I. Glasgow, K.E. Norman, S.D. Bagg, and S.H. Scott, *Assessment of Upper-Limb Sensorimotor Function of Subacute Stroke Patients Using Visually Guided Reaching*, Neurorehabil Neural Repair. 2010 Mar 16. [Epub ahead of print].
- [2] Heart and Stroke Foundation. www.heartandstroke.ca. Date Accessed: March 10, 2007.
- [3] P. H. McCrea¹, and J. J. Eng, *Consequences of increased neuromotor noise for reaching movements in persons with stroke*, Experimental Brain Research, Springer, Berlin/ Heidelberg, pp. 70-77 2004.
- [4] K. Singh, and S.H. Scott, *A motor learning strategy reflects neural circuitry for limb control*, Nature Neuroscience, 6, pp. 399, 2004.
- [5] Gowland, P. Stratford, M. Ward, J. Moreland, W. Torresin, S. Van Hullenaar, J. Sanford, S. Barreca, B. Vanspall and N. Plews. *Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment*. Stroke 1993; 24 (1):58-63.

Online pattern matching in sparse matrices and contact maps

Robert Fraser^{a*}

^aDavid R. Cheriton School of Computer Science,
University of Waterloo,
Waterloo, ON, Canada, N2L 3G1

ABSTRACT

Contact maps are two dimensional abstract representations of protein structures. Some patterns in contact maps correspond to configurations of protein secondary structures. Searching for such patterns may typically use a naïve sliding window approach, and we study techniques which accelerate the searching operations in the online setting, including a restricted search algorithm which operates only on relevant areas of the matrix.

Keywords: String matching, contact maps, protein structure, adaptive analysis

1 INTRODUCTION

The motivation for this problem is protein structure prediction. The contact map is an abstract binary representation of the structure of a protein; creating a contact map from a protein with known structure is a lossy procedure. People have attempted to recover the three dimensional structure of a protein from the contact map [1], but success has been limited. We wish to identify local substructures that can be identified and associated with representative patterns in contact maps, by searching the known body of protein structures. The Protein Data Bank (PDB) contains over 30000 known protein structures at present, so an exhaustive search can be very time consuming. Since the PDB is a dynamic entity, we restrict this discussion to the online setting. The essence of the problem is simple: we have a small rectangular binary pattern which we wish to search for in a database of many large binary patterns.

2 PROTEIN STRUCTURE & CONTACT MAPS

The building blocks of proteins are amino acids, which bond together in a chain to form the structure of the protein. The sequence of these structures is often referred to as the primary structure. Amino acids interact with other amino acids, resulting in secondary structures such as alpha helices. Finally, these secondary structures interact to form the three-dimensional tertiary structure of the protein. The interaction between pairs of alpha helices is the focus of our research.

A contact map can be viewed as an abstract translational and rotational invariant representation of a protein's topology, which captures much of its relevant structural information. A contact map is an $N \times N$ matrix, where N is the number of amino acids in the given protein, and entry C_{ij} in the matrix is a boolean, indicating whether amino acid i is in contact with amino acid j . A threshold distance between atoms is the conventional definition of a contact; values ranging from 7Å to 10Å between C_α atoms are commonly used [2], p.26. It has been shown that regions of contact maps can be used to identify physical properties of pairs of alpha helices [3]. In this case, we often need to search for a small contact pattern (the target) within a large number of source contact maps, such as the entirety of the Protein Data Bank (PDB). A contact map and a refined region corresponding to a pair of alpha helices is shown in Figure 1. Notice that the source contact maps are always square, while the target map is rarely square, since a source map compares the position of every amino acid to every other one, while a target map compares the positions of amino acids in one alpha helix to those in another, and the two alpha helices will rarely be the same length.

3 SEARCHING IN CONTACT MAPS

Consider the pattern shown in Figure 1(b), which corresponds to a pair of alpha helices. To locate pairs of alpha helices with similar properties, the naïve approach is to take this pattern and compare it with each possible position for a match in the source contact map, denoted the 'sliding window' approach. Given a source contact map of size $N \times N$, and a target map of size $I \times J$, the running time is $\theta(N^2 \cdot I \cdot J)$. Note that there is a worst case lower bound of $\Omega(N^2)$ for the matching problem since we desire an exact solution.

*Corresponding author. E-mail: r3fraser@uwaterloo.ca, Telephone: +1(519)885-1211 x35351.

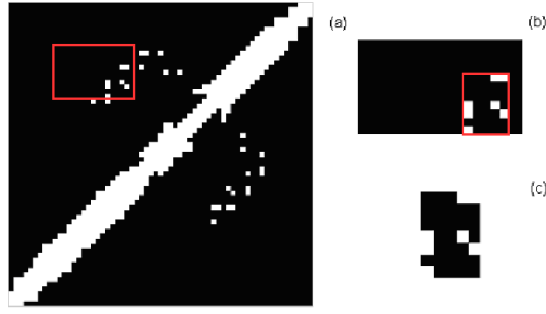


Figure 1. (a) The contact map for protein 1a0a from the Protein Data Bank (PDB). The rectangle indicates the area occupied by two helices, shown in (b). The contact map represents all of the amino acids for one alpha helix along the vertical axis and the other along the horizontal. This has been further refined to the interface area, shown in (c). The contact map interface is found by isolating the smallest rectangle containing all of the contact points from the contact map for the helix pair.

3.1 The Linear Time Algorithm

Bird [4] and Baker [5] (BB) independently discovered linear time algorithms for two-dimensional pattern matching, and here we examine the technique as outlined by Bird. The first step involves searching for the rows of the target in the source pattern. A finite state machine is built that models the transitions representing each row of our target map. We create a trie (or goto function [6]) by moving row by row down the target map, and labelling the states incrementally in the order that we encounter them. The use of a trie rather than a full finite state machine requires the definition of a separate failure function. We find a row of the target by reaching an accepting state, but we still need to search for the other rows above and below the one that we have identified to determine if we have a complete match. Another string matching algorithm is used between the columns of the target and source, where each row of the target is treated as a single symbol. By maintaining an array of size N , we can track the value of the last row found in each column. This algorithm runs in $O(N^2 + I \cdot J)$ [4], and since there are $O(N^2)$ elements to search through in our source contact map and $N^2 \gg I \cdot J$, this is a linear time algorithm in the size of the input.

3.2 Expected Sublinear Pattern Matching

Since we need to look at every element in the source, any algorithm will run in $\Omega(N^2)$ time in the worst case. However, sublinear expected time is possible, as introduced by Baeza-Yates and Régnier [7]. The key insight in their approach is that since we have a pattern with I rows, we really only need to search every I^{th} row of the source for matches. If no target rows are found in row k nor row $k + I$ of the source, then we know that the pattern will not be found in the intervening rows. This algorithm is superlinear in the worst case, $O(N^2 \cdot I)$, but the average case performance is $O(N^2 / I + I \cdot J)$ for randomized data. There are other expected sublinear algorithms, such as that presented by Tarhio [8] which is based on the Boyer-Moore [9] string matching algorithm which searches from right to left. However, the practical performance gains over the others is marginal on the problem sizes we are faced with [8], so we use Baeza-Yates and Régnier’s (BYR) approach as the representative for expected sublinear approaches.

4 ADAPTIVE ANALYSIS OF CONTACT MAP SEARCH

Adaptive analysis is a study of a problem where some properties of the source data can be exploited to achieve both practical and theoretical gains in the complexity of a problem. Perhaps the best known application of adaptive analysis is sorting. Given some sequence of numbers that are to be sorted, it is clear that some sequences are easier to sort than others. This concept leads to the idea of measures of presortedness, which are metrics for quantifying how far from being sorted a sequence is [10]. There are many such measures for sorting, such as the number of inversions, or the maximum distance that an element is from the position that it will occupy in the sorted array.

We wish to apply this type of analysis to the searching of contact maps for pairs of alpha helices. The insight here is clear: there is no point in searching an area of a source contact map if the contacts do not correspond to alpha helices. Our source data for a protein is an array of length N containing the secondary structure for each amino acid and the $N \times N$ contact map. We can now walk along the array and identify regions corresponding to pairs of alpha helices of size greater than or equal to $I \times J$ in time $\theta(N)$. We restrict our search to these regions, and the contact map can be searched as presented in Algorithm 1.

Algorithm 1 is generalized so that any search algorithm can be used as a black box. We implemented each of the algorithms discussed in this paper (naïve, Bird and Baker (BB), and BYR) to determine which was best in this adaptive approach. For our formal analysis, we will first consider the naïve approach. The first loop to identify all

Algorithm 1 The algorithm for Adaptive Contact Map Search. The function takes three arguments: C is the $N \times N$ source contact map, SS is the array of length N giving the secondary structure values for each element, and T is the target contact map of size $I \times J$. We build a set AH , which is a list of the pairs of endpoints of the alpha helices in C .

```

SearchMaps( $C, SS, T$ )
 $AH = \{\}$ 
for  $i = 1, \dots, N$  do
  if  $SS(i) = \text{alpha helix}$  then
     $\text{start} = i$ 
    while  $SS(i) = \text{alpha helix}$  do
       $i = i + 1$ 
    end while
     $\text{end} = i - 1$ 
     $AH = AH \cup \{(\text{start}, \text{end})\}$ 
  end if
end for
for  $i = 1, \dots, |AH|$  do
  for  $j = 1, \dots, |AH|$  do
    if  $i \neq j \ \&\& \ (AH(i).\text{end} - AH(i).\text{start} + 1) \geq I \ \&\& \ (AH(j).\text{end} - AH(j).\text{start} + 1) \geq J$  then
      run search algorithm on  $AH(i) \times AH(j)$  region of  $C$ 
    end if
  end for
end for

```

of the regions of the protein that are in alpha helices takes linear time. The second for loop depends on the number of alpha helices that are found. We define two variables $\xi_I, \xi_J \in [0, 1]$, which represent the fraction of amino acids relevant to the search. For example, suppose that our protein had 100 amino acids, and there are four helices a, b, c, d of length 5, 10, 15, and 20. If our target map is 8×12 , then we only need to search the regions $b \times c, b \times d, c \times d$, and $d \times c$. Further, we subtract I and J from the cost for each pair of helices. We can thus define ξ_I and ξ_J as follows:

$$\xi_{\Xi} = \sum_{i=1}^{|AH|} AH(i).\text{end} - AH(i).\text{start} - \Xi + 1,$$

where $\Xi \in \{I, J\}$ and AH is a list of the alpha helices, such that $AH(i)$ corresponds to the i^{th} alpha helix in the protein. $AH(i)$ stores a pair $\{\text{start}, \text{end}\}$, which are the indices of the first and last amino acids in the i^{th} helix. $|AH|$ is the size of AH (which is the total number of alpha helices in the protein). We set the value $AH(i).\text{end} - AH(i).\text{start} - \Xi = 0$ if it is a negative value for any given i when computing the sum for ξ_{Ξ} . The values for our example are $\xi_I = (0 + 3 + 8 + 13)/100 = 0.24$, $\xi_J = (0 + 0 + 4 + 9)/100 = 0.13$, and $\xi = 0.0312$, where we define $\xi = \xi_I \cdot \xi_J$ to simplify notation (recall that each value is from a different axis, so they are independent). Therefore, the running time of adaptive naïve search is $O(N^2 \cdot \xi \cdot I \cdot J + N)$. If we were searching a contact map that contained zero or one alpha helices, the cost of this search would be $\Theta(N)$.

For the Bird [4] algorithm, we carry over the savings of the linear time algorithm. Their bound was $O(N^2 + I \cdot J)$, and the same arguments above apply to their approach since their algorithm will work just as well on these smaller subregions of the map. Thus, their algorithm takes time $O(N^2 \cdot \xi + I \cdot J + N)$. As indicated by this bound, we expect that the advantages of their approach will be less pronounced in this adaptive scenario. The algorithm of Baeza-Yates and Régnier [7] will have similar worst case performance, and the expected case performance is expressed as $O(N^2 \cdot \xi/I + I \cdot J + N)$.




5 EXPERIMENTAL RESULTS

We searched for matches to the pattern shown in Figure 1 in the PDB using each of the three algorithms discussed in this paper, both using the standard method and the adaptive approach. The results are shown in Table 1. The linear time approach of Bird and Baker is faster than the naïve approach and the expected sublinear BYR algorithm; substantially so in the case of the 7\AA maps. All of the algorithms have fairly equal performance in the adaptive implementations, but the adaptive approach is overall much faster than the original implementations, regardless of which algorithm is used. The reason that the BYR algorithm is not doing as well as might be expected in the original implementation is because of the sparsity of the data. The patterns that we are searching for contain multiple rows

Table 1. The time required by each algorithm to find the chosen pattern, both at the 7Å and 10Å resolution maps.

	Original (min)			Adaptive (min)		
	Naive	BB	BYR	Naive	BB	BYR
7Å	483.8	151.1	391.5	8.1	11.1	10.0
10Å	459.8	148.3	238.5	7.9	8.3	7.0

Table 2. A comparison of the time required by each adaptive algorithm.

	Small 7Å	Small 10Å	No Zeros
Naive (min)	76.2	42.6	17.7
BB (min)	16.6	16.6	18.8
BYR (min)	24.5	16.4	10.3
map			

comprised entirely of zeros, and there are many occurrences of sequences of zeros in the source data sets. Therefore, this algorithm is running close to the superlinear worst case time complexity, $O(N^2 \cdot I)$.

The algorithms have similar performance in the adaptive approaches; each is running at close to linear time. In addition, since the naïve implementation has the sliding window shift as soon as there is a mismatch, the expected running time improves further. To further distinguish the adaptive algorithms, we searched for a small map where the naïve approach should not do as well. Also, we searched for a map with no rows of zeros so that the BYR approach should excel. The results are shown in Table 2, along with images of the target contact maps used in this study. Typically, the maps consist of the interface region of the contact map for the pair plus up to three rows and columns of zeros around the interface if they are present in the original map for the pair (recall that the map for the pair corresponds to Figure 1(b), and the interface is Figure 1(c)). These three extra rows or columns ensure that there is a turn of the helix where there are no further contacts, but impair the performance of the BYR algorithm.

5.0.1 Experimental setup

We used Matlab for implementations with the Bioinformatics toolkit for parsing PDB files. This study could be done using another language to obtain faster search times for each approach, but the relative results should be similar to those obtained here. Also, times reported in this study express the total time used while operating on files, but the time used by Matlab for opening and closing individual PDB files was not included.

6 CONCLUSIONS AND FUTURE WORK

We have presented several algorithms for searching for a small contact map pattern in numerous large source contact maps for exact matches in an online setting. Each of them provides better performance than the naïve algorithm, increasing the tractability of these search problems, as was observed through their implementation and application. Contact maps are typically sparse, as are the patterns that we are searching for, which results in poor performance for the naïve algorithm. The speed-up expected from the BYR algorithm is often negated due to sparsity. Based on these observations, the Bird approach is usually the best choice.

This paper presents several new ideas. This is the first known adaptive analysis to be performed on the two dimensional string matching problem. Although it is application specific, the general technique may be applied in other areas given some domain-specific knowledge analogous to our secondary structure information. We presented experiments suggesting that the BB algorithm is best when searching for patterns with rows of zeros in sparse matrices. In the adaptive setting, the best algorithm was dependent on the number of rows containing zeros in the target pattern. When performing searches where one is padding the interface with zeros, the BB approach is best, while the BYR algorithm is better for searches for interfaces with no blank rows.

There are several other refinements that may be carried out to further improve the speed of searching. One is to perform the matching on compressed data. Due to the sparsity of the data, contact maps seem ideal for run-length encoding schemes, and such techniques allow better performance than online approaches. Further, it is possible that the database could be indexed, allowing instantaneous searches.

ACKNOWLEDGMENTS

The author wishes to thank Janice Glasgow for the inspiration for this project and Alejandro Salinger and Arash Farzan for their insights and discussions. Finally, the detailed and high quality feedback received from the anonymous reviewers is greatly appreciated.

References

- [1] M. Vendruscolo, E. Kussell, and E. Domany, “Recovery of protein structure from contact maps,” *Folding and Design* **2**(5), pp. 295–306, 1997.
- [2] R. Fraser, “A tale of two helices: A study of alpha helix pair conformations in three-dimensional space,” 2006.
- [3] R. Fraser and J. Glasgow, “A demonstration of clustering in protein contact maps for alpha helix pairs,” in *Proc. Int’l Conf. on Adapt. and Nat. Comp. Algs (ICANNGA)*, LNCS **4431**, pp. 758–766, 2007.
- [4] R. Bird, “Two dimensional pattern matching,” *Inf. Proc. Lett.* **6**(5), pp. 168–170, 1977.
- [5] T. Baker, “A technique for extending rapid exact string matching to arrays of more than one dimension,” *SIAM J. on Comp.* **7**, pp. 533–541, 1978.
- [6] A. Aho and M. Corasick, “Efficient string matching: an aid to bibliographic search,” *CACM* **18**(6), pp. 333–340, 1975.
- [7] R. Baeza-Yates and M. Régner, “Fast two dimensional pattern matching,” *Inf. Proc. Lett.* **45**, pp. 51–57, 1993.
- [8] J. Tarhio, “A sublinear algorithm for two-dimensional string matching,” *Patt. Rec. Lett.* **17**, pp. 833–838, 1996.
- [9] R. Boyer and S. Moore, “A fast string matching algorithm,” *CACM* **20**, pp. 762–772, 1977.
- [10] V. Estivill-Castro and D. Wood, “A survey of adaptive sorting algorithms,” *ACM Computing Surveys* **24**(4), pp. 441–476, 1992.

Modulation transfer function of amorphous selenium digital x-ray detectors

Yuan Fang,^{a*}, Nicholas Allec^a, Ling Guo^b, and Karim S. Karim^a

^aDept. of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada ;

^bDept. of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada

ABSTRACT

We evaluate the modulation transfer function (MTF) of amorphous selenium (a-Se) digital x-ray detectors. This study includes the effects of generation and reabsorption of characteristic x-rays, which can significantly degrade the detector MTF. Monte Carlo (MC) methods are used for simulation of spatial dose distribution, and the detector MTF is computed by Henkel Transform. We consider mammography and chest radiography x-ray energies and detector thicknesses in this study. Incident photon energies of 12 keV, 13 keV, 50 keV and 100 keV are simulated for 150 μm , 300 μm , and 1 mm thick a-Se x-ray detectors. Significant MTF degradation is observed at incident photon energies higher than the a-Se K-edge, due to generation and reabsorption of characteristic x-rays.

Keywords: Amorphous selenium, modulation transfer function, direct detection, characteristic x-rays

1 INTRODUCTION

X-ray imaging is commonly used by physicians to view the internal organs and structures of human body and diagnosis of bone fractures and suspicious lesions for cancer. Digital radiography can provide many advantages over traditional film-based radiography, such as dose reduction and convenience of image processing. Amorphous Selenium (a-Se) is an excellent photoconductive material, with an effective atomic number for high x-ray absorption, low dark current, and can be uniformly deposited on large areas, making it the only commercially available x-ray photoconductive material for direct conversion digital x-ray detectors. For this study, a-Se detector is considered for medical imaging modalities in mammography and chest radiography.

One of the important metrics for evaluating an x-ray detector is the modulation transfer function (MTF). This metric is a function of spatial frequency and takes into account effects such as the reabsorption of characteristic x-rays. [?] The MTF allows us to better determine the maximum resolution of the detector, thus it is of great importance. We look at the MTF for a-Se and we show the effect of incident photon energy and thickness on the detector MTF.

2 METHODS

This section describes in detail the simulation methods of our study.

2.1 Modulation transform function

This subsection includes more detailed equations for MTF calculations. The definition of the MTF include include the x-ray detector MTF (reabsorption of characteristic x-rays) and the aperture effect due to the pixel size. [?]

$$\text{MTF}_{\text{pre}}(k) = \text{MTF}_x(k) \times |\text{sinc}(\pi a_{\text{del}} k)| \quad (1)$$

The detector MTF can be computed by Henkel transform, and the point spread function (PSF). For this study, the PSF is constructed from simulation results. [?]

$$\text{MTF}_x = H[p(r)] = 2\pi \int_0^\infty p(r) J_0(2\pi k r) r dr \quad (2)$$

The aperture effect has a sinc squared dependence on the pixel width, and is the reciprocal of the volume under the squared MTF, give by [?]

$$a_e = \left[2\pi \int_0^\infty \text{MTF}^2(k) k dk \right]^{-1} \quad (3)$$

*Corresponding author. E-mail: y4fang@uwaterloo.ca, Telephone: +1(519)635-9320.

The calculation of the PSF from simulation results will be covered in the following subsections.

2.2 Monte Carlo code

The latest version of the Photon Electron Shower (PENELOPE) Monte Carlo code [?] was used to simulate the photon and electron interactions within a-Se photoconductor. PENELOPE allows for simulation of detailed transport of photon and electrons, and mean free path as a function of the particle's incident energy is shown in Figure 1. Our simulation model takes into account all possible photon and electron interaction mechanisms in the diagnostic energy range. The incident photon can interact inside the a-Se detector by Rayleigh scattering, Compton scattering, photoelectric absorption and pair production, where only Compton and photoelectric absorption lead to creation of secondary electrons. This secondary electron can interaction in the detector material by elastic scattering, inelastic scattering and Bremsstrahlung. Only inelastic collisions lead to energy loss and dose deposition that is required for MTF calculations.

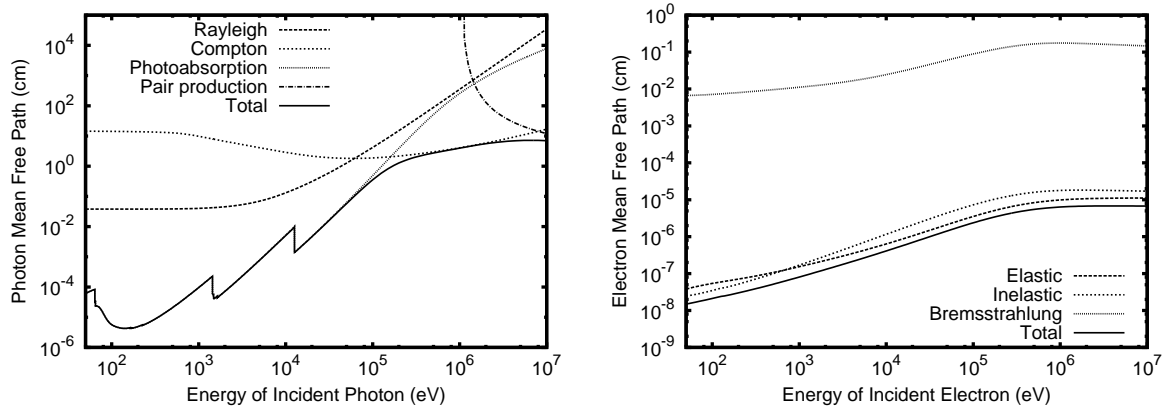


Figure 1. (a) Mean free path of photons in a-Se as a function of incident photon energy. (b) Mean free path of electrons in a-Se as a function of electron energy

2.3 Detector geometry

The detector is modeled by a single layer of a-Se. The detector model as shown in Figure 2 consist of an mono-energetic photon beam of x-ray incident perpendicularly on the center of a cylindrical detector of 10 cm in diameter. The a-Se detector is further subdivided radially with an equal spacing of $0.5 \mu\text{m}$ each. Figure 2 illustrate the modeled detector geometry.

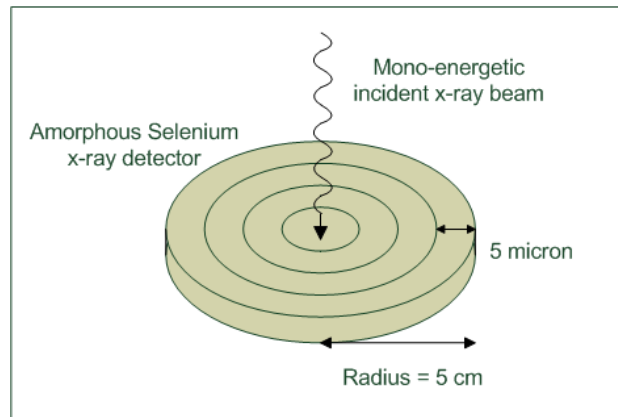


Figure 2. Detector geometry in cylindrical coordinates modeled by Monte Carlo simulation.

For calculation purposes, the density used for a-Se is 4.5 g/cm^3 . MC simulations were used to generate the dose distribution profile $d(r)$, which is the energy deposition inside the detector material with respective radial distribution.

The PSF is constructed from this dose profile using the relation below. Where the dose is normalized to the unit area.

$$p(r) = \frac{d(r)}{2\pi \int d(r) r dr} \quad (4)$$

3 RESULTS

This section shows the simulation results of detector MTF for mammography and chest radiography applications, not including the aperture dependence. Figure 3(a) shows the MTF for 150 μm thick a-Se detector for mammography application, with incident photon energy of 12, 13, 50 and 100 keV. The MTF is highest for the 12 keV case because of low incident energy lead to minimal lateral energy spreading and it is below the a-Se K-edge of 12.6 keV. For energies above this K-edge, for example 13 keV, a sharp drop of MTF is observed. This is due to the reabsorption of characteristic x-rays produced above the 12.6 keV K-edge. As incident energy increases, the more energy is deposited locally by electrons and the MTF is considerably improved. However, the MTF degrades eventually at energies above 100 keV.

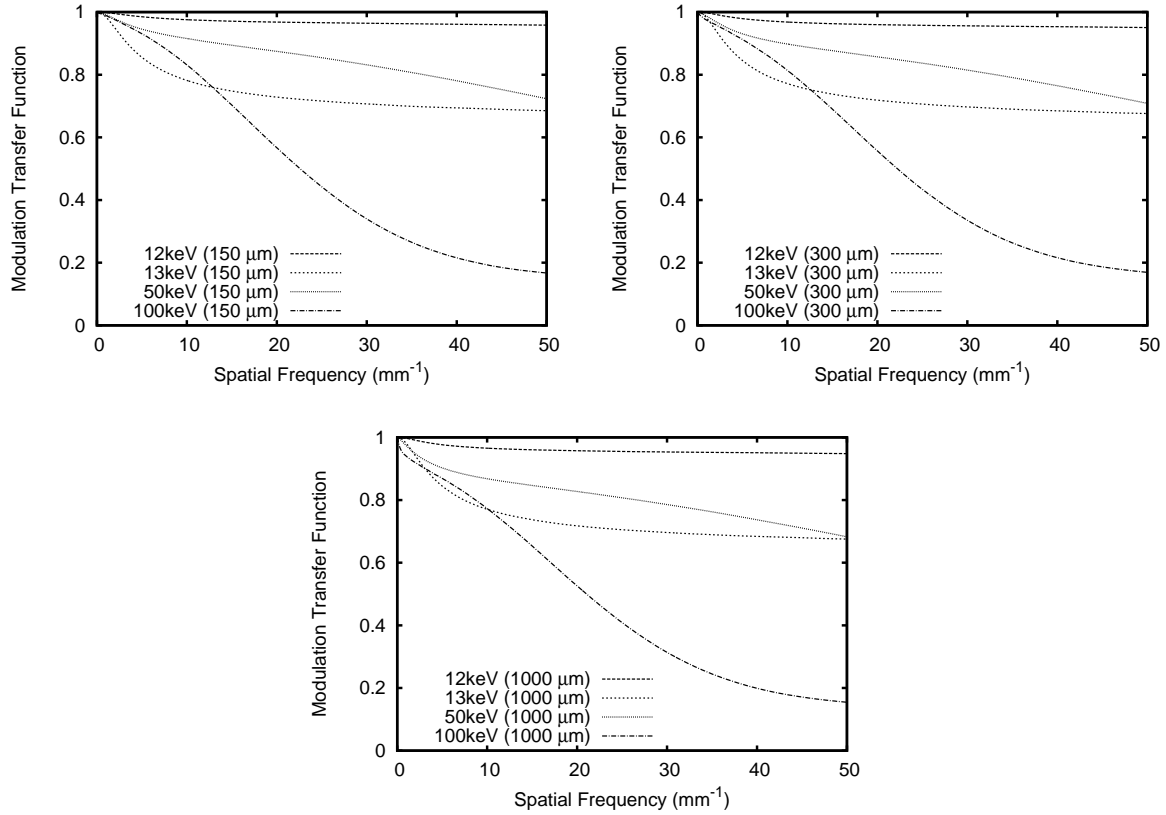


Figure 3. (a) MTF for 150 μm thick a-Se detector for mammography. (b) MTF for 300 μm a-Se detector. (c) MTF for 1000 μm a-Se detector for chest radiography.

Figure 3(b) and (c) shows similar trend of MTF for 300 μm thick a-Se detector and 1 mm thick detector commercially used for chest radiography applications. Similar MTF degradation are observed due to reabsorption of characteristic x-ray photons above the K-edge of a-Se.

4 CONCLUSIONS AND FUTURE WORK

We have shown the MTF in a-Se digital x-ray detectors, and demonstrated that the reabsorption of characteristic x-rays are significant in degradation of MTF. The detector MTF is simulated as a function of incident photon energy and detector thickness, and this shows a good indication of the resolution achievable by a-Se detectors for mammography and chest radiography applications. For future work, the aperture effect can be included in the simulation model, and further improve accuracy of calculation.

ACKNOWLEDGMENTS

This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC) and Waterloo Institute of Nanotechnology.

References

- [1] G. Hajdok, J. J. Battista, and I. A. Cunningham, “Fundamental x-ray interaction limits in diagnostic imaging detectors: spatial resolution,” *Med. Phys.* **35**, pp. 3180–3193, June 2008.
- [2] H. Barrett and W. Swindel, *Radiological Imaging*, Academics, New York, 1981.
- [3] G. Hajdok, J. J. Battista, and I. A. Cunningham, “Fundamental limitations imposed by x-ray interactions on the modulation transfer function of existing x-ray detectors,” in *Medical Imaging 2003: Physics of Medical Imaging*, Feb. 2003.
- [4] F. Salvat, J. Fernandez-Varea, and J. Sempau, “PENELOPE-2006: A code system for Monte Carlo simulation of electron and photon transport,” in *Workshop proceedings, Organisation for economic co-operation and development*, July 2006.

Using Dynamic Bayesian Networks to analyze genetic data

Zhen Wang^{a*} and Andrew Wong^b

^aSchool of Computing, Queen's University, Kingston, Canada;

^bSchool of Computing, Queen's University, Kingston, Canada

ABSTRACT

In System biology, many statistical approaches are used to analyze gene expression data and infer gene regulatory network. Dynamic Bayesian Network(DBN) is a well-known method that has produced promising results when used to analyze time-series data. Here, we adopt a Dynamic Bayesian Network algorithm with Markov Chain Monte Carlo (DBmcmc) developed by Dirk Husmeier to investigate gene expression data. We evaluated the algorithm's performance on a synthetic yeast time-series data and then applied it to the rat CNS development temporal data [1] to reverse engineer the gene network. The resulting network is then validated against previous studies on the same data set as well as evidence from biological databases and literature. The interactions documented in biological literature show that DBmcmc was able to correctly identify subnetworks of interacting genes that were involved in the same biological pathways .

Keywords: Dynamic Bayesian Network, Markov Chain Monte Carlo, Gene Networks, Structure Learning

1 INTRODUCTION

With the development of microarray technology, determining gene networks from microarray expression data has become a main research focus in the post-genomic era. A genetic network attempts to model the interactions between genes to discover casual relationships. There are many different methods for determining gene networks. One promising approach is Dynamic Bayesian Networks (DBNs). The statistical properties of DBNs allow it to estimate the relationships among genes objectively. Furthermore, DBN allows scientists to incorporate prior knowledge into the data analysis algorithm. Also, DBN can handle time series data and therefore can reflect causability. However, the complexity for its structure learning is exponential to the number of variables and is considered a NP-hard problem [2]. Normally, it is too computationally expensive to use DBN for gene expression data due to the large number of genes involved. Therefore, search heuristics that reduces the number of potential graph structures must be used to generate the DBN. Many search heuristics exist for learning the structure of DBN. Here we adopt DBN with Markov Chain Monte Carlo (DBmcmc) developed by Dirk Husmeier to analyze our time-series gene expression data.

The main goal of this project, therefore, is to discover biologically relevant gene regulatory network using DBmcmc. Next section gives a brief introduction about DBmcmc. The datasets and their gene networks that are produced by DBmcmc will be presented in the Results section. Also, the learned structure will be validated using evidence from biological databases (KEGG Pathway, GO, NCBI) and literature.

2 Methods

A Bayesian Network(BN) is a directed acyclic graph, G . The graphical structure G includes nodes which denote variables and edges between the nodes represent their conditional dependencies.

The learning of a Bayesian Network is basically searching for the graph with the highest posterior probability given data. Based on the Bayesian Formula, given the gene expression data D , the probability of G is:

$$P(G|D) = \frac{P(G,D)}{P(D)} = \frac{P(D|G)P(G)}{P(D)}. \quad (1)$$

Theoretically, with a sufficiently large set of data, the graph structure that exactly captures all dependencies in the distribution will receive a higher probability than all other graphs, as proposed by Friedman and Yakhini [3].

*Zhen Wang,E-mail: zhenw@cs.queensu.ca, Telephone: +1(613)533-6000 ext 74201

In a Dynamic Bayesian Networks model, we observe gene expression values at different time points. The assumption of Dynamic Bayesian Network is that an event at time t is only influenced by event at time $t - 1$, and the conditional probability is defined as

$$P(X_t|X_{t-1}) = \prod_i P(X_t^i|Pa^G(X_t^i)). \quad (2)$$

Dynamic Bayesian Network is essentially a two-slice Bayesian Network that captures temporal relationships. The nodes in the first slice do not have parameters associated with them, and the ones in the second slice have an associated conditional probability distribution which is given by (2). As is shown in Figure 2, DBN can reflect recurrent networks well.

As mentioned previously, the number of possible network structures is exponential to the number of nodes. Therefore, searching through all possible structures to find the optimal network is impractical. Therefore, a search heuristic must be employed in order to generate the resulting network. In our study, we apply Husmeier's method mentioned in [4]. The algorithm, Markov chain Monte Carlo simulation, starts with random nodes as root nodes and generates a sequence of networks [5] based on the Metropolis-Hastings acceptance criterion (MHAC) to determine the final network [6]. Given an initial network G_{old} , the algorithm generates a new network G_{new} by either adding, removing, or reversing an edge. This network is then scored using the MHAC, which is a function based on the posterior probability of the new network given data, P , and the probability of the new network given the old network, Q . This MHAC can be expressed as follows:

$$P = \min\{1, \frac{P(G_{new})|D}{P(G_{new}|D)} \times \frac{Q(G_{old}|G_{new})}{Q(G_{new}|G_{old})}\}. \quad (3)$$

where the posterior probability of the network given data is determined using Bayesian Scoring (2). At the end of the MCMC simulation, the algorithm samples a user-defined number (e.g. 20,000) of networks from the generated sequence. The network with the highest posterior probability at the end of the simulation is then chosen as the final network. The algorithm also includes a burn-in step where an initial number (e.g. 20,000) of networks are discarded from the chain before sampling. This is because the initial networks are not stable and therefore, are not reliable.

The DBN with Markov Chain Monte Carlo algorithm is implemented in MATLAB using the DBmcmc toolbox written by Dirk Husmeier, which invokes subroutines of the BNT toolbox given by Murphy, both of which are publicly available online.

3 Results

3.1 Synthetic Data

To evaluate the performance of this method, we first apply it to a synthetic yeast time series simulation data given by Husmeier. This simulated time series is binary in nature due to requirements of DBmcmc. The dataset is generated from a true yeast cell cycle network G_0 from Friedman [7] and its real network is shown in Figure 1(a). We used the dataset to produce a Dynamic Bayesian Network using DBmcmc. The result is then compared with the true network to evaluate the performance of DBmcmc.



(a) True Bayesian network G_0 . It is of a sub network of the yeast cell cycle, taken from Friedman [7]; 38 unconnected nodes were included, 50 nodes totally.

(b) Calculated network after adding noise. The solid line between ACE2 and RNR3 is false discovered and dotted lines mean the missing edges.

Figure 1. Synthetic dataset.

We first applied DBmcmc on the synthetic data. If we choose the posterior probability of 0.5 as a threshold to cut edges, two false edges are discovered. If we set the threshold as 0.8, the true network will be recovered by the algorithm. This shows that DBmcmc can effectively recover the underlying gene network from data. Furthermore, we also checked the consistency of the algorithm by adding noise to the data and re-evaluating its performance. We randomly changed 10% of the data by flipping the binary expression values and re-ran DBmcmc on the modified data. The resulting network from the noisy data, with the threshold 0.8 is shown in Figure 1(b).

This result shows that even with 10% noise in the data, the algorithm is still able to detect most of the true edges. Moreover, since noise is added randomly, the algorithm will generate different networks every time. On average, across repeated simulations (250 times), the algorithm was able to recover about 80% of the true edges.

3.2 Real rat CNS data

3.2.1 Data preprocessing

In our study, we used a subset with 65 genes from the original data generated by Wen *et al* [1]. Since the expression values in this dataset are continuous, it had to be preprocessed before DBmcmc can be applied. This involved the discretization of the dataset into three categories: under-expressed, normal, and over-expressed.

First, in order to make gene expressions comparable, we normalized data and did cubic spline interpolation to have more time points. To discretize the data, we calculated the mean μ_i and standard deviation σ_i of each gene $Gene_i$, $i = 1, 2, \dots, 65$ correspondingly. For $Gene_i$, if the expression value is between $\mu_i - \alpha\sigma_i$ and $\mu_i + \alpha\sigma_i$, we set its value as 0 (normal), less than $\mu_i - \alpha\sigma_i$ is -1 (under), and the rest will be set as 1 (over). By experimenting with different thresholds, we empirically determined that $\alpha = 1.0$ gave the optimal discretization. By setting $\alpha = 1.0$, 68% of the gene expression would be classified as normally expressed and are only over-expressed or under-expressed 32% of the time.

3.2.2 Gene Network

The DBmcmc simulation outputs a $n \times n$ matrix that shows the interactions between genes in the DBN as a conditional probability. The algorithm begins with random root nodes, different simulations produce different results. Therefore, to generate our gene interaction network, we conducted 5 simulations and selected the edges that appeared consistently in the different outputs. To do this, we took the sum of the five $n \times n$ interaction matrices, and applied a threshold to remove interactions where the sum of its probability was less than two. On average, this meant that an interaction have to appear be in at least 3 networks with probability greater than 0.7 to be included. Using this strategy, the gene network is produced, 56 genes are included with 94 interactions identified.

3.2.3 Validation

Since the network as a whole is hard to analyze due to its size and connectivity, we focused on identifying and validating selected subnetworks.

In order to extract subnetworks that may be significant, we first grouped the genes into groups based on the functional categories highlighted by Wen *et al.* in 1998 [1]. Then, based on these functional categories, we looked for subnetworks where the majority of the participating genes belong to the same functional group. Based on this heuristic, 4 potential subnetworks were selected.

SubNetwork(SN)	Functional Groups
SN1	Neurotransmitter Metabolizing Enzymes (GAD-), Glutamate Receptor (mGluR-)
SN2	Heparin-binding Growth Factors (-GF)
SN3	Acetylcholine Receptors (nAChR-)
SN4	Serotonin Receptors(5HT-), GABA-A Receptors (GR-)

In SN1 shown in Figure(2(a)), there were many interactions between different forms of GAD and mGluR genes. Evidence from literature supports these interactions as GAD seems to be involved with GABA synthesis [8] and mGluR has an inhibitory effect on GABA [9]. Aside from interactions between these two genes, PDGF and IP3R were also shown as parents of mGluR and GAD. These interaction are also relevant because according to the KEGG database, both PDGF and IP3R are involved in a Calcium Signalling Pathway and Calcium release causes the release of glutamate, which binds to the receptor mGluR [10].

In the second subnetwork, SN2, the gene GAP43 is identified as the parents of many Growth Factors (GF). This interaction resonates with the fact that GAP43 is a regulator of PDGF, IGF, and FGF. [11], [12], [13] This subnetwork also show an interaction between EGF and bFGF, which are genes that are related neuron growth [14].

Our third network also identified interactions between genes that reflect the underlying biology. In SN3, BDNF is identified as a parent of nAChR, which correlates with the findings from Massey *et al.* [15] that BDNF increases

ACKNOWLEDGMENTS

This paper is a course project done under the instruction and help from Dr. Parvin Mousavi.

References

- [1] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi, "Large-scale temporal gene expression mapping of central nervous system development," *Proc Natl Acad Sci* **6**, pp. 334–339, 1998.
- [2] D. M. Chickering, "Learning bayesian networks is np-complete," in *Learning from Data: Artificial Intelligence and Statistics*, D. Fisher and H. Lenz, eds., pp. 121–130, Springer, NY, 1996.
- [3] N. Friedman and Z. Yakhini, *On the sample complexity of learning Bayesian networks*, 1996.
- [4] D. Husmeier, "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks," *Bioinformatics* **19**, pp. 2271–2282, 2003.
- [5] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *Amer. Statist.* **49**, pp. 327–335, 1985.
- [6] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika* **57**, pp. 97–109, 1970.
- [7] N. Friedman, M. Linial, and I. Nachman, *Using Bayesian Networks to Analyze Expression Data*, 2000.
- [8] H. Wu, Y. Jin, C. Buddhala, G. Osterhaus, E. Cohen, H. Jin, J. Wei, K. Davis, K. Obata, and J. Y. Wu, "Role of glutamate decarboxylase (gad) isoform, gad65, in gaba synthesis and transport into synaptic vesicles-evidence from gad65-knockout mice studies," *Brain Research* **1154**, pp. 80–83, 2007.
- [9] A. van den Pol, X. Gao, P. Patrylo, P. Ghosh, and K. Obrietan, "Glutamate inhibits gaba excitatory activity in developing neurons," *J Neurosci.* **18**, pp. 10749–10761, 1998.
- [10] J. Petravicz and K. McCarthy, "Loss of ip3 receptor-dependent ca²⁺ increases in hippocampal astrocytes does not affect baseline ca1 pyramidal neuron synaptic activity," *J Neurosci.* **28**, pp. 4967–4973, 2008.
- [11] R. Curtis, R. Hardy, R. Reynolds, B. Spruce, and G. Wilkin, "Down-regulation of gap-43 during oligodendrocyte development and lack of expression by astrocytes in vivo: Implications for macroglial differentiation," *Eur J Neurosci.* **3**, pp. 876–886, 1991.
- [12] F. Piehl, H. Hammarberg, T. Hokfelt, and S. Cullheim, "Regulatory effects of trophic factors on expression and distribution of cgrp and gap-43 in rat motoneurons," *J Neurosci* **51**, pp. 1–14, 1998.
- [13] L. Mohiuddin, P. Fernyhough, , and D. Tomlinson, "Acidic fibroblast growth factor enhances neurite outgrowth and stimulates expression of gap-43 and α 1 α -tubulin in cultured neurones from adult rat dorsal root ganglia," *Neuroscience Letters* **215**, pp. 111–114, 1996.
- [14] K. Baldauf and K. Reymann, "Influence of egf/bfgf treatment on proliferation, early neurogenesis and infarct volume after transient focal ischemia," *Brain Research* **1056**, pp. 158–167, 2005.
- [15] K. Massey, W. Zago, and D. Berg, "Bdnf up-regulates α 7 nicotinic acetylcholine receptor levels on subpopulations of hippocampal interneurons," *Mol Cell Neurosci.* **33**, pp. 381–388, 2006.
- [16] J. Feng, X. Cai, J. Zhao, and Z. Yan, "Serotonin receptors modulate gaba(a) receptor channels through activation of anchored protein kinase c in prefrontal cortical neurons.," *J Neurosci* **21**, pp. 6502–6511, 2001.
- [17] F. Kawahara, H. Saito, and H. Katsuki, "Inhibition by 5-ht₇ receptor stimulation of gabaa receptor-activated current in cultured rat suprachiasmatic neurones," *J Physiol.* **478**, pp. 67–73, 1994.
- [18] P. Haeseleer, "Linear modeling of mrna expression levels during cns development and injury," *Pacific Symposium on Biocomputing* **4**, pp. 41–52, 1999.
- [19] T. Yoshinori, K. SunYong, B. Hideo, I. Seiya, T. Kousuke, K. Satoru, and M. Satoru, "Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection," *Bioinformatics* **14**, pp. 227–236, 2003.
- [20] N. Friedman and D. Koller, "Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks," *Machine Learning* **50**, pp. 95–125, 2003.

MeSHOP: MeSH Over-Representation Profiles for Summarising Biomedical Literature

Warren A. Cheung^{ad1}, BF Francis Ouellette^b and Wyeth W. Wasserman^{cd}

^aBioinformatics Program, UBC, Vancouver, BC, CANADA;

^bOntario Institute for Cancer Research, Toronto, ON, CANADA

^cMedical Genetics, UBC, Vancouver, BC, CANADA

^dCentre for Molecular Medicine and Therapeutics, Vancouver, BC, CANADA

ABSTRACT

The ever-expanding library of biomedical literature provides researchers with a unique dilemma of summarising the increasingly-unwieldy existing publications on a specific topic. Statistical analysis of Medical Subject Heading terms provides a quantitative profile of the relevant biomedical terms, through the analysis of the primary literature for the topic of interest. These profiles are filtered to remove redundancies and the analysis modified to highlight the most relevant biomedical terms. As well, terms in profiles can be inferred by profile similarity, as demonstrated by the novel association of disease terms to gene profiles via gene profile to disease profile similarity. These novel associations are shown to be predictive of gene-disease association in literature. Further analysis shows the extent and age of the literature associated to a gene are correlated to future association in disease-related publications.

Keywords: gene profiles, disease profiles, MeSH term profiles, indirect association, overrepresentation profiles

¹ Corresponding author. E-mail: wcheung@cmmmt.ubc.ca, Telephone: + 1(604)875-2345 ext. 7947

Studying the Effect of Pulsatile Compulsory Blood Flow in Models of Stenotic Coronary Artery by Application of the Fluid-Structure Interaction

Alireza Hashemifard^{a1} and Nasser Fatourae^{b2}

^aBiomedical Engineering Faculty, Science and Research branch of IAU, Tehran, IRAN;

^bBiological Fluid Dynamics Research Laboratory, Biomedical Engineering Faculty, Amirkabir University of Technology (Tehran Polytechnic), Tehran, IRAN 15914;

ABSTRACT

The stenosis geometry, and its severity, has important influence on recirculation length, and distribution of macromolecules concentration, such as Low Density Lipoproteins (LDL) in arteries. Here a research has been conducted to study the effects of the artery's elasticity feature and the stenosis severity on the blood flow pattern of coronary artery. An elastic model of the artery with a stenosis is created which is used as the test bench throughout this research. The model is equipped with a particular inlet and is used to investigate the effect of various stenosis severities and detailed flow pattern phenomenon. The results have indicated that the presence of stenosis causes serious variations in the flow pattern and provides a suitable situation for production of multiplicity of stenosis and increase of the stenosis length in the coronary artery. The increase in the pressure due to the presence of stenosis is noticeable and produces stress in artery's wall. The elasticity of the artery has also significant impact on the flow patterns and pressure variations. The stenosis creates vorticities which their existence exacerbates the severity of the stenosis in the first place. One of the observations made is the increase in blood pressure before stenosis which leads to the artery dilation and its shear stress increase. This stress is even more increased, in turn, by escalation of the stenosis severity. The trapping of the blood particles in the vorticities, which can be considered as a sign for the creation of clot or consecutive stenosis, has been demonstrated by particle tracing of the fluid in this simulation. The change of the stenosis entrance condition from the blood compulsory flow to the pressure-dependent flow, seen in this simulation, is an important observation that helps to understand the decrease in the amount of the blood flow in the artery.

Keywords: Blood flow, Stenosis, Dilation, Compression, Fluid-Structure interaction, Compliance

¹Corresponding author: Alireza Hashemifard, E-mail: hashemifard@bmedoc.com

²Nasser Fatourae, PhD., Phone: (+98-21) 64542368, Fax: (+98-21) 66468186, E-mail: nasser@aut.ac.ir, WWW: <http://bme.aut.ac.ir/~nasser>

Aspects of Beta Amyloid Aggregation and Its Interaction with Acetylcholine Neurotransmitters and Alzheimers Disease

Ibrahim Mustafa, Pu Chen, Ali Elkamel

Chemical Engineering Department, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

Alzheimers disease is the most popular form of cholinergic diseases with over 25 million people suffering from it worldwide. β -amyloid aggregates destroy brain cells, causing severe problems with memory, thinking and behavior and eventually leading to death. The interaction between β -amyloid aggregates and Acetylcholine (ACh) is predicted for various operation ranges and specific initial conditions of feed parameters. Characterization of the response of ACh production, and other parameters to β -amyloid aggregation is made. Greater understanding of physiological behaviour in terms of sudden disturbances in parameters by sensitivity analysis is achieved.

Image Segmentation Using Varying Ellipses

Farnoud Kazemzadeh, Thomas M. Haylock, and Arsen R. Hajian

Department of Systems Design Engineering, University of Waterloo, 200 University Avenue
West, Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

Highly accurate segmentation techniques are needed in the field of medical imaging. Segmenting medical images is difficult due to varying contrast and the high level of noise inherent in many medical imaging modalities. A segmentation routine based on growing ellipses is shown to be able to segment an object from background in these images. The algorithm increases the size of each ellipse until a threshold is met, indicating the edge of the object. The object's center position is initialized by the algorithm operator and a suitable threshold is found iteratively and interactively. Using standard image processing test images, results show as high as 98.5% successful segmentation. Success is gauged by comparing results to a manually segmented image and false positive segmentation tends to be low. Suitable images for ellipse segmentation are symmetric and well enclosed, which are typical characteristics found in a medical image.

Segmentation of Carotid Artery from Three-dimensional Ultrasound Images Using Active Contours

E. Ukwatta^{ab}, J. Awad^a, A. D. Ward^a, A. Krasinski^a and A. Fenster^{abc}

^aImaging Research Laboratories, Robarts Research Institute, ^bThe Department of Medical Biophysics, ^cBiomedical Engineering Graduate Program, The University of Western Ontario, London, Ontario, Canada

ABSTRACT

In the research setting, three-dimensional ultrasound (3D US) imaging is being developed to provide highly sensitive, robust and precise measurements of carotid artery atherosclerosis. To quantitatively evaluate carotid artery atherosclerosis, accurate measures of plaque volume, surface morphology and composition are required and these dictate that rapid, precise and accurate segmentation of the carotid vessel walls and lumina from 3D US images. Manual segmentation is extremely time consuming and operator-dependent. Therefore, the objective of this study is to develop and validate a semi-automatic segmentation algorithm for delineating the vessel wall and lumen of carotid arteries for patients with asymptomatic carotid stenosis. Carotid arteries are extremely challenging to segment using image information alone due to the presence of plaque, poor definition of the vessel boundaries, intensity heterogeneity, image speckle and shadowing. Therefore, we combine various image cues with domain knowledge of the vessel geometry and some user interaction into the segmentation framework. We adopted an energy minimization approach based on the level sets method to segment the vessel wall and lumen of carotid arteries using edge-based and region-based objective functions respectively. The proposed segmentation method was evaluated with respect to manually outlined boundaries using several similarity measures on 60 2D US slices from ten patients. Our method yielded Dice coefficients of $91\% \pm 0.05\%$, $90 \pm 0.06\%$ and mean absolute distance errors (MAD) of 0.32 ± 0.13 mm, 0.27 ± 0.14 mm for vessel wall and lumen segmentations, respectively. The realization of semi-automated methods will accelerate the translation of 3D US to real time clinical research and clinical care.

Keywords: VWV, segmentation, level sets, 3D US, carotid plaque

Quantification of prostate deformation due to needle insertion during TRUS-guided biopsy

Tharindu De Silva^{a,b}, Aaron Fenster^{a,b}, Jagath Samarabandu^a, Aaron D. Ward^b

^aGraduate Program in Biomedical Engineering, The University of Western Ontario, London, ON, Canada.

^bImaging Research Laboratories, Robarts Research Institute, London, Canada.

ABSTRACT

Prostate cancer is one of the most common cancers among men, second only to skin cancer. Prostate biopsy is the clinical standard for the diagnosis of prostate cancer, and technologies for 3D guidance to targets and recording of biopsy locations are promising approaches to reducing the need for repeated biopsies. In order to biopsy the smallest clinically significant tumors with 95% confidence, the RMS error of the biopsy system should be less than 2.5mm. There can be multiple potential sources of error that can cause the actual target biopsy location to be different from the expected target in such systems, including: (1) tolerances in the design and construction of mechanical needle guidance systems, (2) errors in imaging and calibration to the needle guidance systems, (3) patient and prostate motion and deformation during the procedure due to interaction with the TRUS probe and discomfort during biopsy, (4) prostate deformation due to slow biopsy needle insertion in preparation for biopsy gun firing, and (5) prostate deformation due to rapid biopsy needle insertion after firing the biopsy gun. There is previous research in measurement of and correction for the first three sources of error. However, to the best of our knowledge, prostate deformation due to needle insertion through the rectal wall and biopsy gun firing has not yet been quantified in the context of TRUS-guided prostate biopsy. In our study, we use image-based non-rigid registration to quantify prostate deformation during needle insertion and biopsy gun firing, in order to provide information useful to the overall assessment of a TRUS-guided biopsy system's expected targeting error. We recorded mean tissue displacements of up to 0.4 mm, accounting for 16% of the clinically-motivated maximum desired RMS error of guidance system.

Cross-laboratory comparison of human post-mortem brain expression profiling data

Meeta Mistry¹, Kelsey Hamer², Paul Pavlidis^{2,3}

¹ CIHR/MSFHR Bioinformatics Training Program, Univ. Of British Columbia, Vancouver, BC, Canada;

² Centre for High-throughput Biology;

³ Department of Psychiatry

ABSTRACT

Expression profiling of post-mortem human brain tissue has been widely used to study molecular changes associated with neuropsychiatric diseases. Changes in expression associated with factors such as age, or gender can mask or complicate the detection of expression patterns attributable to disease. In the current study we have performed a large meta-analysis of genome-wide expression studies of normal human cortex to more fully catalogue the effects of age, gender, post-mortem interval and brain pH, yielding a “meta-signature” of gene expression changes for each factor. We used the Gemma system (<http://chibi.ubc.ca/Gemma>) and other bioinformatics tools to combine datasets across multiple studies to identify expression signatures with increased sensitivity. In addition to the inherent value of the meta-signatures, our results provide critical information for future studies of disease effects.

COMPUTER-CONTROLLED DYNAMIC HUMAN GASTROINTESTINAL MODEL: IN VITRO EXPLORATION OF THE HUMAN GUT MICROBIOTA

Rodes L.^{a1}, Tomaro-Duchesneau C.^{a1}, Coussa-Charley M.^{a1}, Martoni C.^{a1},
Bhathena J.^{a1}, Prakash S.^{a1}

^aBiomedical Technology and Cell Therapy Research Laboratory, Department of Biomedical Engineering, Faculty of Medicine, McGill University, 13775 University Street, Montreal, Quebec, H3A 2B4, CANADA

ABSTRACT

In the recent past, there has been a growing interest in the potential benefits of the trillions of bacteria that compose the human gut microbiota. This complex ecological system has long been known for its necessary role in the production of vitamin K, in the transformation of bile acids and in the absorption of ions. New evidence demonstrates that the composition of the gut microbiota participates in the prevention and development of various disorders, including inflammatory bowel diseases, colon cancer and hypercholesterolemia. In vitro models to simulate the human gastrointestinal (GI) tract allow for the functional and bacteriological exploration of the human gut microbiota. Few continuous complex flora batch systems have been developed. Here, we describe the computer-controlled dynamic human GI model that is utilized in our laboratory. It consists of a succession of 5 bioreactors representing, respectively, the stomach, the small intestine, and the ascending, transverse and descending colon of the human GI tract. The microbial ecosystem is sustained by automatically feeding a food solution to the first vessel. Temperature, pH, volume, residence time and pH of each reactor are computer-controlled. The fermentation vessels are maintained in anaerobic conditions. Therefore, each reactor harvests the microflora of a different region of the human GI tract. This is the only dynamic human GI model available in North America. It offers an excellent and reliable system to investigate the human gut microbiota ecology, activity and stability. One of the main applications of this computer-controlled dynamic human GI model is for the screening of probiotic bacteria, to assess their characteristics as an oral delivery system such as, effectiveness and robustness, and, to determine the effects of these biotherapeutic formulations on the human gut microbiota.

Keywords: Gastrointestinal tract, in vitro model, continuous batch system, gut microbiota, probiotic bacteria, oral delivery system, biotherapeutics

¹ Corresponding author. E-mail: satya.prakash@mcgill.ca.

A real-time biomechanical analysis method for multifocal breast cancer assessment

Shadi Shavakh^{a,1}, Aaron Fenster^{b,c} and Abbas Samani^{a,b,c}

^aDepartment of Electrical & computer Engineering, University of Western Ontario, London, ON, Canada;

^bDepartment of Medical Biophysics, university of Western Ontario, London, ON, Canada;

^cImaging Research Laboratories, Robarts research Institute, London, ON, Canada

ABSTRACT

According to the Canadian Cancer Society in 2009, breast cancer is the second most commonly diagnosed cancer among Canadian women. In most diagnosed cases there is a single tumour, however, recent research has found that up to 60 % of breast cancer is multifocal. Early diagnosis and classification of breast cancer is a critical step in choosing an appropriate treatment plan. In our laboratory, a novel ultrasound elastography method has been developed. This method is capable of imaging absolute Young's modulus (YM) of breast tumours in real-time fashion. In this technique, we acquire the tissue strain field and surface force data as an input for tissue YM reconstruction. We obtain the surface force data to calculate stress distribution throughout the region of interest by using a Statistical Finite Element Method (SFEM), which was recently developed in our laboratory for fast stress analysis. This presentation demonstrates our novel ultrasound elastography technique for imaging multifocal breast cancers. The YM reconstruction technique is iterative and involves stress calculation in each iteration. We use the SFEM technique for stress calculation paving the way for real-time YM reconstruction. The fundamental idea of statistical shape model is that there is a high degree of shape and stress distribution similarity between specific organs such as the breast under given load. To develop the breast SFEM we used pre-processed data obtained from FE analysis of a large number of similar objects in a statistical shape model framework. For force data acquisition, a system consisting of two load cells was developed to measure forces on the breast surface. This data is used as input for calculating stress distribution. This stress distribution is combined with the measured strain data to update the YM distribution. Numerical and tissue mimicking phantoms were tested with this elastography system and very encouraging results were obtained..

Keywords: Breast cancer, Ultrasound, Elastography, Real-time, Multifocal tumour

¹ Shadi Shavakh. E-mail: sshavakh@uwo.ca, Telephone: +1(519)702-5856.

Fractal Time-Series Analysis of Reaching Motion in Stroke Affected Arms

Tanny F. King¹, Kathrin Tyryshkin¹ and Janice I. Glasgow¹

¹Queen's University, 93 University Ave., Kingston, Canada

ABSTRACT

A stroke is caused by an interruption of blood flow to the brain, and its persisting effects depend on the location and damage incurred in the brain. One possible effect is an impairment of motor function in the upper limbs. Assessment of the impairment is necessary for effective rehabilitation and recovery. Although current clinical assessments such as CMSA and FMA are reliable, these tests may be subjective and imprecise as they do not examine movement precisely. Analyses of time series collected by robotic devices may provide more accurate and less biased assessments. Since many physiological time series are fractal in nature, it was hypothesized that the time series of a reaching motion was fractal and that the same fractal properties would no longer be preserved in stroke affected subjects. The time series data was collected during a center outreach task using a KINARM device. The series were first analyzed using power spectral density (PSD) analysis which revealed a fractional Brownian motion ($\beta \approx 2.02$) for both control and stroke groups. Based on these results, the time series was then subjected to bridge detrended scaled window variance analysis and Higuchi's fractal dimension analysis to estimate the Hurst coefficient (H) and the fractal dimension (D) respectively. The results indicated $H \approx 0.95$ and $D \approx 1.02$ in all test groups. Regardless that each method inferred fractal structure individually, the estimated parameters do not agree with one another according to literature and thus the reaching motion may not be truly classified as a fractal. However, the estimated parameters were significantly different at $\alpha=0.05$ between the stroke affected arms and the controls. This may suggest new ways to differentiate between stroke and control subjects, and may eventually lead to development of useful objective scores for diagnosing the impairment with further research.

Topology-aware closest point cortical thickness

Eli Gibson ^{a*} and M. Faisal Beg ^{b†}

^aImaging, Robarts Research Institute, London, Canada;

^bSchool of Engineering Science, Simon Fraser University, Burnaby, Canada;

ABSTRACT

Evidence for regional cortical thickness changes associated with disease onset has been found for many neurodegenerative diseases, including Alzheimer’s disease, schizophrenia, AIDS, autism, and ADHD. The automated computation and analysis of cortical thickness facilitates the detection of these changes, enabling early and differential diagnosis of these diseases and assisting medical researchers in identifying causal and curative factors.

A plethora of methods to measure the cortical thickness have been described, the majority of which are based on either closest point thickness, which measures the minimal distance between points on the inner and outer cortical surfaces, or coupled surface thickness, which defines a mesh on one cortical surface, propagates the mesh to the opposing surface, and measures the distance between corresponding points. Closest point thicknesses are mathematically simple and intuitive; however, measurements on the highly convoluted cortex suffer from errors, due to incorrectly representing anatomy. Coupled surface methods avoid these errors by explicitly incorporating anatomical boundaries, but lose the conceptual simplicity of closest point thickness.

We first describe idiosyncratic anatomy that leads to errors in closest point methods, and demonstrate the prevalence of such anatomy in 340 brains from the OASIS database [1], showing idiosyncratic anatomy in all brains, with an average of 460 instances per brain. We then present a novel hybrid method that yields the advantages of both techniques, by creating a coupled surface method that preserves anatomy, whose correspondences are based on minimal distances as in the closest point methods. Coupled surfaces are generated via the spherical registration of coordinate functions [2] of the pial and white cortical surfaces. We demonstrate the improved performance of this method in idiosyncratic anatomy, and assess the robustness of the method using repeated measurements relative to closest point methods.

Keywords: cortical thickness, closest point, coupled surfaces, idiosyncratic anatomy

References

- [1] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults,” *J Cogn Neurosci* **19**, pp. 1498–1507, Sep 2007.
- [2] E. Gibson, A. R. Khan, and M. F. Beg, “A combined surface and volumetric registration (SAVOR) framework to study cortical biomarkers and volumetric imaging data,” *Medical Image Computing and Computer-Assisted Intervention* **5761**, pp. 713–720, 2009.

*Corresponding author. E-mail: egibson@robarts.ca, Telephone: +1(226)919-6786.

†E-mail: mfbeg@ensc.sfu.ca, Telephone: +1(778)782-5696, <http://autobrainmapping.com/>

Design and Implementation of a 3D Ultrasound System for Image Guided Liver Interventions

Hamid R. Sadeghi-Neshat^{a,b,1}, Shi Sherebrin^b, Lori Gardi^b, Aaron Fenster^{a,b,c}

^a Biomedical Engineering Graduate Program, The University of Western Ontario, London, ON

^b Imaging Research Laboratories, Robarts Research Institute, London, ON

^c Department of Medical Biophysics, The University of Western Ontario, London, ON

ABSTRACT

Several minimally invasive, image-guided procedures have been developed for the detection and local treatment of hepatocellular carcinoma and hepatic metastases. One limitation of these methods is the difficulty in locating the preoperative planning data within intraoperative images for accurate guidance and placement of the instrument. 2-D ultrasound imaging is the most commonly used intraoperative guidance method. The fusion of multiple 3-D ultrasound scans, taken from different orientations and at different times, can enhance interventional procedures accuracy and may be used to quantify response to therapy.

The first objective of this study was to design a new device for scanning and reconstruction of 3D ultrasound scenes from 2-D image planes. Acquired images are visualized with an interactive software user interface designed in our group to provide image guidance during the procedure. Main source of error in minimally invasive liver procedures is organ deformation and motion due to respiration. To reduce this error, an automatic registration approach is under investigation. The proposed registration technique is used, first to align pre-operative interventional plans (usually obtained from CT or MR images) with live ultrasound images, and then to update the plan on the ultrasound images acquired in different phases of the intervention. Our technique is a combination of landmark and intensity based registration approaches.

The first prototype of the 3D ultrasound liver scanner was manufactured and used to acquire images from tissue-mimicking phantoms and two volunteers. Preliminary results of registration between ultrasound image sets acquired during different respiration cycles and with differing scanner orientations were used to assess the algorithm's accuracy. Target registration error (TRE) and fiducial localization error (FLE) are measured and reported to estimate the registration error.

In order to track the instrument/lesions in real-time and to match them with preoperative planning data, in the next phase of the project we will design new tools to guide the instrument to the desired target accurately.

Keywords: Image-guided interventions, liver cancer, 3D ultrasound, image registration

¹ Corresponding author. E-mail: hneshat@imaging.robarts.ca

Applying Normal Mode Analysis to the Conserved Patterns of Cytochrome C.

En-Shiun Annie Lee

University of Waterloo, 200 University Avenue West, Waterloo, Canada;

ABSTRACT

Normal mode analysis is a classical mechanics technique for studying harmonic potential wells analytically. This simulation technique studies large-scale internal dynamics of proteins. In this paper, normal mode analysis is used to study the movement of Cytochrome C. in order understand its sequence conservation. To understand the conservation of the protein sequence and its patterns, the dynamic movement of the protein must be understood. This paper set out to discover the relationship between protein sequence conservation and protein movement by using protein alignment data to study the sequence conservation and normal mode analysis to study protein movement. The structure is as follows: the first section is a detail description of the differential calculus and physics behind normal mode analysis; the second section presents the Cytochrome C. family and its conserved pattern, the analysis of a case study on the protein 3CP5, and a framework for automating normal mode analysis is described. The results from the case study found that the highly conserved “C**CH” pattern demonstrates rigid non-movement, a correlation between sequence conservation and dynamic movement, and conserved patterns have less movement in general.

Keywords: Normal Mode Analysis, Significant Pattern, Invariant Site, Cytochrome C.

Monte Carlo simulation of amorphous Selenium digital x-ray detectors: spatial dependence of energy absorption

Yuan Fang^{a1}, Nicholas Allec^a, and Karim S. Karim^a

^aDept. of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada

ABSTRACT

Amorphous Selenium (a-Se) digital x-ray detectors can be used for various medical imaging modalities including mammography, and fluoroscopy. One major performance limitation of a-Se x-ray detectors is the spatial blurring caused by reabsorption of fluorescent photons. Fluorescent photons are produced when an incident x-ray photon interact within the detector material. Initially, a high-energy secondary electron is ejected from the inner shell of the atom, followed by relaxations of the atom that may lead to creation of fluorescent photons. Compared to the secondary electrons, fluorescent photons have high mean free path, and may be reabsorbed within the material at a large distance away from the interaction cite, thus causing significant spatial blurring in the projected x-ray image. The effect of fluorescent reabsorption is dependent on the incident x-ray energy spectrum, and thickness of the detector, while the effect of spatial blurring is dependent on the pixel size. For example, an increase in the detector thickness would lead to an increase in the probability of reabsorption; while the effect of spatial blurring can be reduced by increasing the pixel size, and decreasing the inherent detector spatial resolution. Hence, there is a fundamental limitation in spatial performance, and the detector design can be optimized to minimize the effect of spatial blurring caused by fluorescent photons. With the use of an accurate and benchmarked Monte Carlo simulation package, PENELOPE, we examine the spatial distribution of energy deposition within a-Se digital x-ray detectors for mammography and fluoroscopy applications. The choice of thickness and pixel dimensions take into account the spatial dependence of the energy absorption. We present results on both lateral and vertical (i.e. depth) energy deposition within the detector for mammography and fluoroscopy applications using typical diagnostic x-ray spectra.

Keywords: Monte Carlo, Amorphous selenium, digital x-ray detection, energy absorption

¹ Corresponding author. E-mail: y4fang@uwaterloo.ca, Telephone: +1(519)635-9320.

Reconstruction of Needle Tracts from Fluoroscopy in Prostate Brachytherapy

Lauren E. Gordon^a, Ehsan Dehghan^a, Septimiu E. Salducean^b and Gabor Fichtinger^a

^aQueen's University, Kingston, Canada

^bUniversity of British Columbia, Vancouver, Canada

ABSTRACT

Purpose: In prostate brachytherapy, a surgeon implants radioactive seeds using needles to irradiate cancer while sparing healthy tissue. Seed positions can be reconstructed from fluoroscopic images taken before and after the procedure to assess quality. Reconstructing needles from these seed positions can help us to better understand needle bending, tissue deformation and seed migration. It can also be used for needle-needle ultrasound-fluoroscopy registration for intraoperative dose assessment and planning.

Methods: Needle reconstruction can be formulated as an assignment problem, where each seed is matched to the seed next to it along the length of a needle. Since seed assignment costs can be represented as a bipartite graph, the global lowest-cost solution can be found by the Hungarian Algorithm in polynomial time. Our method uses the Hungarian Algorithm to find the best seed assignments, and then uses these assignments to trace all of the needles.

Results: Prostates were simulated as spheres, with needles curved toward the central axis. Seeds were placed on needles with a random perturbation of up to 1/3 of the needle spacing. With these simulated seed positions, our method resulted in over 95% of seeds being assigned to the correct neighbour and correct needle in less than 5s at 2.00GHz and 1GB RAM. Furthermore, the algorithm was tested on one clinical data set, with a success rate of over 98% in less than 3s. The combination of speed and accuracy suggests that the algorithm may be used for intraoperative applications.

Keywords: Prostate cancer, brachytherapy, fluoroscopy, Hungarian Algorithm

Notes

Notes

Notes

Notes