BIOT-04
Proceedings of the First Biotechnology and
Bioinformatics Symposium

Sponsored by
The University of Colorado at Colorado Springs
The Colorado Institute for Technology Transfer
and Implementation

September 24, 2004
Colorado Springs, Colorado.
Preface

This volume contains the papers accepted for oral presentation at BIOT-04, The 1st Symposium on Bioinformatics and Biotechnology: A Community and Academic Forum held on the 24th of September in Colorado Springs, Colorado.

Biotechnology is starting to have a major impact on society. Innovative drugs have come to the marketplace attempting to cure and control diseases that afflict us. Products of food biotechnology have created seeds and plants that produce tastier and healthier food with fewer environmental impacts. With the completion of sequencing of the human genome, bioinformatics and the related field of computational biology are also playing very important roles in science, technology, and society. Huge amounts of information are being produced continually by gene-sequencing experiments, combinatorial chemistry synthesis, gene-expression investigations and proteomics studies. Research and development in bioinformatics attempt to distill useful information from mountains of data. Bioinformatics is a result of convergence of biology and computing. Impacts of bioinformatics are being felt in many areas such as the discovery of new genes responsible for diseases, design and development of new drugs, and the study of origins of life in terms of phylogenetic trees.

In organizing this symposium, we have several objectives. First, we want to create a platform where researchers in biotechnology and bioinformatics can share their research with other academicians, individuals from industry, as well as interested individuals from the community at large. We want to organize a small, integrated and high-quality symposium where individuals such as biologists, computer scientists, mathematicians, lawyers, and social scientists can come together and learn from each other. Second, we want to create a tradition whereby we can continue to have such a symposium every year in a beautiful part of the country where we want to foster research and development in biotechnology and bioinformatics. For this purpose, we have already tentatively scheduled an extended two-day symposium next year in August. There are no biotechnology and bioinformatics symposiums or conferences of note in the Rocky Mountain area and we would like our symposium to grow and become nationally recognized. We also want to reach out internationally.

We are happy at the results of our first effort. We were able to put together a Program Committee with about 40 members from around the nation including the University of Pennsylvania, Columbia University, Rice University, Louisiana State, and Georgia State University. We were also lucky to get several individuals from local as well as national companies such as Monsanto to be on the Program Committee. We also have international members from Wales and India on the Program Committee. We received 35 papers from around the US, Canada, India and Russia. Due to the fact that we have only one day for the symposium, and we did not want to break into parallel sessions so that attendees are exposed to interdisciplinary presentations, we were able to accept only 11 out 35 papers for presentation, giving us an acceptance rate of about 30%. We accepted 11 additional papers for poster presentation. We think it is an auspicious beginning.

I would like to thank all the members of the Program Committee, every author who submitted papers and abstracts, everyone who is attending this symposium, and everyone who has a poster or is giving an oral presentation. It is wonderful to have you all here. We are also grateful to El Pomar Foundation for allowing us to hold this symposium in the beautiful premises of the Penrose House, the Colorado Institute for Technology Transfer and Implementation (CITTI) for helping organize this symposium, and the University of Colorado at Colorado Springs for providing us with the computing, office and other facilities. We are also very thankful to the Colorado Bioscience Association for their sponsorship. Finally, my thanks are to Symanta Saikia of Wichita, Kansas, who designed the cover for us. The cover contains repeated patterns containing the ankh, the original cross from Egypt representing both physical and eternal life. It also contains the West African symbol sese woruban that represents transformations in life. This is a common symbol found on cloth and walls, in pottery and logos all over Ghana.

Welcome to BIOT-04 and our fair City of Colorado Springs. Enjoy your stay and come back next year!

Jugal Kalita (jkalita@uccs.edu, http://www.cs.uccs.edu/~kalita)
On Behalf of BIOT-04 Program Committee
Department of Computer Science, University of Colorado at Colorado Springs, 1420 Austin Bluffs Parkway
Colorado Springs, CO 80917 USA

All contents © UCCS, CITTI
# TABLE OF CONTENTS

## ORAL PRESENTATIONS

**Cell-based Screening of Human Breast Cancer-Associated Gene Collection Identifies Novel Functional Activities** .......................................................... 1  
Lisa Hines, Abigail Witt, Yanhui Hu, Ru Gunawardane, Steven Isakoff, Nicole Collins, Joshua LaBaer and Joan Brugge

**Clinically Relevant Medical Image Compression** .......... A  
Matthew J. Zukoski, Terrance Boul, Tije Iyriboz

**LP3 - The LEA Protein Prediction Pipeline** .......................................................... 11  
Martin Gollery, Barrett Abel, John Cushman

**Jumpstarting Phylogenetic Analysis** .......................................................... 17  
Mark Clement, Quinn Snell, Kevin Seppi, Keith Crandall

**Recognizing TATA Promoters Based on Discriminating Frequency Analysis of Neighborhood Tuples** .......................................................... 25  
Raja Loganathanraj, M E Karim, Arun Lakhota

**Computational Modeling Simulation of the Immune System** .......................................................... 30  
Jugal Kalita, Kaushal Chandrashekar, Ankur Deshmukh, Reena Hans, Priyadarshini Selvam

**Can new technology impact the Pharmaceutical Industries innovation shortfalls?** .......................................................... 36  
Layne Los, Virgil Gary Tye

**Growing the Family Tree: The Power of DNA in Reconstructing Family Relationships** .......................................................... 42  
Luke Hutchison, Natalie M. Myres, Scott Woodward

**Software modeling of the Complement System and its role in Immune Response** .......................................................... 50  
Aparna Srikanthaswamy, Jugal Kalita, Kaushal Chandrashekar, Karen Newell, Patricia Giclas

**HALO - A Multifunctional Proliferation Assay for Basic, Clinical and Drug Development Research** .......................................................... 56  
Ivan N. Rich

**Two Experiments in Biological Term Annotation using Classification Methods** .......................................................... 61  
Sittichai Jampaniarn, Vlado Keselj, Nick Cercone

## POSTERS

**Unsupervised Evolution of Probabilistic Neural Network for Clustering Gene Expression Data** .......................................................... 66  
Daron Anderson

**In Silica Characterization of Rad26, a Fission Yeast DNA Damage Checkpoint Protein** .......................................................... 68
Kristyn Baker, Jugal Kalita, Tom Wolkow

**Rapid and Efficient Dissemination of Biological Data with Interactive Keys: The Southern Rocky Mountain Interactive Flora** ................................................................. 71
Jeffrey Wayne Brasher and Neil Snow

**CoCoLysis: A Web-Accessible Coiled-Coil Protein Database with Analysis Tools** ................................................................. 73
David Brinkmann, Sai Nandoor, Jugal Kalita, Brian Tripet, Robert Hodges

**Mathematical Model for the Human Arterial System** 79
Robert Carlson, Radu C. Cascaval

**Character Based Phylogeny Inference in the Grid Computing Environment** ................................................................. 81
Zhijie Guan

**A Web-Based Tool for Modeling and Simulation of Ion Channels** ................................................................. 83
Mona Soliman Habib, Jugal Kalita

**Novel Soybean Antibiotics** ................................................................. 89
Tamara J McEwen, Daniel M Zurek

**Applied Sampling Methods for Estimating the Total Number of SNPs in Human Chromosome 19** ................................................................. 90
Leigh Quiogue, Christopher Chee

**Analysis of Nasal Epithelium Gene Expression in Cystic Fibrosis** ................................................................. 92
Thanh Richidi, Frank Accurso, Krzysztof Cios

**AGENDA**

- 7:30 am Check-In and Continental Breakfast
- 8:00 am - 9:00 am Speaker - Ronald B. Moss, MD, PhD
- 9:00 am - 9:30 am Break/Poster Review
- 9:30 am - 11:15 am Session I Presentations

**Cell-based Screening of Human Breast Cancer-Associated Gene Collection Identifies Novel Functional Activities**
Lisa Hines, Abigail Witt, Yunhui Hu, Ru Gunawardane, Steven Isakoff, Nicole Collins, Joshua LaBaer, Joan Brugge

**Clinically Relevant Medical Image Compression**
Matthew J. Zukoski, Terrance Boul, Tunc Iyriboz

**LP3 - The LEA Protein Prediction Pipeline**
Martin Gollery, Barrett Abel, John Cushman

**Jumpstarting Phylogenetic Analysis**
Mark Clement, Quinn Snell, Kevin Seppi, Keith Crandall

**Recognizing TATA Promoters Based on Discriminating Frequency Analysis of Neighborhood Tuples**
Raja Loganathanraj, M E Karim

**Computational Modeling Simulation of the Immune System**
Jugal Kalita, Kaushal Chandrashekar, Ankur Deshmukh, Reena Hans, Priyadarshini Selvam

- 11:15 am - 1:15 pm Lunch with Poster Session and Speakers - Susan P. James, PhD & Susan C. Trapp, PhD
- 1:15 pm - 3:15 pm Session II Presentations
Can new technology impact the Pharmaceutical Industries innovation shortfalls?
Layne Los, Virgil Gary Tye

Growing the Family Tree: The Power of DNA in Reconstructing Family Relationships
Luke Hutchison, Natalie M. Myres, Scott Woodward

Software modeling of the Complement System and its role in Immune Response
Aparna Srikantaswamy, Jugal Kalita, Kaushal Chandrashekar, Karen Newell, Patricia Giclas

HALO - A Multifunctional Proliferation Assay for Basic, Clinical and Drug Development Research
Ivan N. Rich

Two Experiments in Biological Term Annotation using Classification Methods
Sittichai Jampoonjamarn, Vlado Keselj, Nick Cercone

- 3:15 pm - 3:30 pm  Break/Poster Review
- 3:30 pm - 4:30 pm  Keynote Speaker - Richard Goldsby, PhD
- 4:30 pm - 5:30 pm  Poster Review/Networking

INVITED SPEAKERS

- **Richard A. Goldsby, PhD**, Professor of Biology & John Woodruff Chair Lecturer, Amherst College; Adjunct Professor, Department of Veterinary & Animal Sciences, Univ. of Massachusetts
- **Ronald B. Moss, MD, PhD**, Vice President of Medical Affairs, Telos Pharmaceuticals; Editor-in-Chief, Journal of Immune Based Therapies and Vaccines
- **Susan P. James, PhD**, Associate Professor, Mechanical Engineering, Biomedical Engineering Program, Colorado State University
- **Susan C. Trapp, PhD**, Computational Bioscience Program Liaison, University of Colorado Health Science Center

PROGRAM COMMITTEE

- **Colin Askill**, Wales National Health Service Trust, Swansea, Wales (Clinical Oncology)
- **Ian Askill**, Aspire Biotech, Inc., Colorado Springs, CO (Wound Healing, Biomaterials)
- **Sandra Berry-Lowe**, Department of Biology, University of Colorado at Colorado Springs
- **Meredith D. Betterton**, Department of Applied Mathematics, University of Colorado at Boulder
- **Dhruba Kr Bhattacharyya**, Tezpur University, Assam, India (Machine learning, Data Mining)
- **Terrance Boult**, Department of Computer Science, University of Colorado, Colorado Springs, CO
- **Cory Butz**, Department of Computer Science, University of Regina, Saskatchewan, Canada (Probabilistic Reasoning)
- **Radu Cascaval**, Department of Mathematics, University of Colorado, Colorado Springs, CO (Mathematical Modeling in Physiology)
- **Nick Cercone**, Department of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada (Knowledge Representation, Expert Systems, Database Learning, Knowledge Discovery)
- **Mark Clement**, Department of Computer Science, Brigham Young University, Provo, UT (Phylogenetic Trees, Parallel Computing, Supercomputers)
- **Krzysztof Cios**, Department of Computer Science, University of Colorado at Denver, CO (Bioinformatics, Data Mining, Machine Learning, Neural Networks)
- **Kanna Das**, Columbia Genome Center, Columbia University, New York, NY (Tissue Culture, Embryonic Stem Cell, Knockout Techniques)
- **Pradip Das**, Monsanto Company, St. Louis, MO (Analytical Sciences, High-thruput Technology)
Sanjoy Das, Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS (Evolutionary Algorithms, Computational Genomics)

Emilia Evtcheva, Department of Biomedical Engineering and Department of Physiology and Biophysics, State University of New York at Stony Brook, NY (Cardiac Cell Function and Engineering)

Paul Fishwick, Department of Computer Science, University of Florida, Gainesville (Modeling, Simulation, Visualization)

Patricia Giclas, National Jewish Medical and Research Center, Denver, CO (Immunology, Complement System)

Lisa Hines, Institute of Proteomics, Harvard University, Cambridge, MA (Functional Genomics and Molecular Epidemiology)

Robert Hodges, Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, Denver, CO (Biomolecular Structure, Peptide Chemistry)

Jean Jacob, Louisiana State University Eye Center, Baton Rouge, LA (Ophthalmic Therapies and Devices)

Jugal Kalita, Department of Computer Science, University of Colorado, Colorado Springs, CO (Bioinformatics, Artificial Intelligence)

Sue Kane, Aspire Biotech, Inc., Colorado Springs, CO (GLP (Glucagon-like peptide) systems, pharmaceutical and agrochemical analysis)

James Mattoo, Department of Biology, University of Colorado, Colorado Springs, CO

Eric Neufeld, Department of Computational Science, University of Saskatchewan, Saskatoon, Canada (Artificial Intelligence, Visualization)

M. Karen Newell, Associate Professor and Merkert Endowed Chair of Biology, University of Colorado, Colorado Springs, CO (Immune-Mediated Cell Death; Tumor Immunology; Cellular Metabolism)

Thomas O'Donnell, Siemens Corporate Research, Princeton NJ (Biomedical Imaging and Modeling)

Yi Pan, Georgia State University, Athens GA (Bioinformatics, Parallel and Distributed Computing)

Saurav Pathak, Department of Physics, University of Pennsylvania, Philadelphia, PA (Multi-variate Methods, Machine Learning)

Ann Prewett, Replication Medical Expertise, New Brunswick, NJ (Bone Regeneration)

Ivan Rich, HemoGenix, Inc., Colorado Springs, CO

Rinaldo Schinazi, Department of Mathematics, University of Colorado, Colorado Springs, CO

Ambuj Singh, Department of Computer Science, University of California at Santa Barbara, CA (Bioinformatics, Databases, Distributed Computing)

Daedra Studniarz, Colorado Institute for Technology Transfer and Implementation, Colorado Springs, CO (Technology Transfer)

Devika Subramanian, Department of Computer Science, Rice University, Houston, TX (Artificial Intelligence, Adaptive Systems, Modeling Visualmotor Tasks)

Virgil Gary Tye, NextGen Pharma Technologies, Inc., Colorado Springs, CO

Tom Wolkow, University of Colorado at Colorado Springs, CO (Cell Biology)

Cheng-Zhong Xu, Wayne State University, Detroit, MI (Distributed and Parallel Computer Systems, Medical Engineering)

Xiaobo Zhou, University of Colorado at Colorado Springs, CO (Distributed Computing, Information Fusion)

Matthew Zukoski, Wilkes University, Wilkes-Barre, PA (Biomedical Imaging and Modeling)
Cell-Based Screening of Human Breast Cancer-Associated Gene Collection Identifies Novel Functional Activities

(Extended Abstract)

Lisa Hines and Abigail Witt, Yanhui Hu, Ru Gunawardane, Steven Isakoff, Nicole Collins, Joshua LaBaer and Joan Brugge
Harvard Institute of Proteomics and Harvard Medical School, Department of Cell Biology
Boston, MA 02115
Lisa.Hines@uchsc.edu, joshua_labaer@hms.harvard.edu, joan_brugge@hms.harvard.edu

ABSTRACT

Functional genomic approaches that comprehensively evaluate the biological activities of human genes may provide novel insight into disease etiology. To systematically investigate the genetic events involved in breast carcinogenesis, we created the Breast Cancer 1000 (BC1000), a sequence-validated cDNA collection of 1000 unique breast cancer-associated genes. Cell-based, functional assays were conducted on a subset of the BC1000 collection to identify proteins that induce phenotypic changes characteristic of both the premalignant and malignant stages of breast cancer. With this comprehensive approach, we identified potentially novel functional activities for known breast cancer-associated genes. To our knowledge, the BC1000 represents the first publicly available sequence-validated human disease gene collection.

1. INTRODUCTION

The human adult breast is an organized, branch-like structure comprised of ducts and lobes. Lobes are composed of multiple individual units termed acini, which consist of a hollow lumen and a single layer of polarized epithelial cells that are surrounded by myoepithelial cells and a basement membrane. Disruption of this well-ordered structure is a prerequisite for carcinoma formation. Increased proliferation of epithelial cells, loss of acinar structure and filling of luminal space are common characteristics of the premalignant stages of breast cancer, such as ductal carcinoma in situ (DCIS). The transformation from premalignant stages to the potentially lethal metastatic tumor occurs when cells lose anchorage dependence, invade the basement membrane and metastasize. The genetic events that trigger the conversion from a normal breast to the premalignant stages of breast cancer and the subsequent progression to malignancy are not well understood.

It is likely that many molecular factors contribute to the progressive stages of breast cancer. To systematically investigate the genetic events involved in breast carcinogenesis, we conducted a functional screen to identify proteins that induce phenotypic changes characteristic of both the premalignant and malignant stages. This strategy involved two phases: assembling a ready-to-use clone collection of candidate genes and screening this collection in three cell-based phenotypic assays to identify genes that induce cancer-like phenotypes.

In the first phase, we assembled the Breast Cancer 1000 (BC1000), a sequence-validated cDNA collection of 1000 breast cancer-related genes ranging from well-studied oncogenes to the less conspicuous breast cancer-associated genes. In the second phase, a subset of genes in the BC1000 collection was introduced into an immortalized mammary epithelial cells (MCF-10As) and tested for their ability to induce phenotypic characteristics associated with premalignancy and malignancy, specifically cellular proliferation, disruption of acinar morphogenesis, and cell migratory behavior independently or cooperatively with activated ErbB2. Through this comprehensive approach, we identified potentially novel functional activities for known breast cancer-associated genes.

2. METHODS

2.1 SELECTION OF THE GENES

The majority of genes included in the BC1000 were selected by MedGene, a literature-mining
software application that searches all titles and abstracts in the Medline database to identify genes co-cited with a particular disease and utilizes statistical methods to rank the relative strengths of these gene-disease relationships based on the frequency of total citation and co-citation [1].

2.2 CONSTRUCTION OF THE BC1000

The coding regions of the genes were captured by PCR amplification of either first strand cDNA or plasmids from the NIH Mammalian Gene Collection (MGC). Using high-throughput methods, the BC1000 gene collection was assembled into the Creator™ system (BD Biosciences), which allows the transfer of genes using Cre-lox site specific recombination. Clones were produced in two formats: one with a universal stop codon (TAG) to allow expression of native protein and one where the stop codon has been replaced with Leucine (TTG) to allow carboxyl-terminal peptide fusions. Amino-terminal peptide fusions can be added to both forms. All constructs were validated with full-length sequencing.

2.3 CELL-BASED ASSAYS

The generation of MCF-10A cells expressing the chimeric receptor p75.B2 has been described [2]. cDNAs (with a STOP codon) were transferred into the pBabe-puro vector [3]. MCF-10As and MCF-10As with the inducibly activated ErbB2 receptor (10A.B2) were plated and infected with VSV G pseudotyped retroviruses expressing the BC1000 collection as described [4].

EGF-independent proliferation assay. EGF-independent proliferation in two-dimensional cultures were performed by plating $3 \times 10^4$ MCF-10A cells in 6.0cm-diameter dishes in either the presence or the absence of EGF (5ng ml$^{-1}$). Genes were scored as a hit if cellular proliferation was observed to be more than with vector alone.

Morphogenesis assay. Three-dimensional assays using MCF-10A cells were performed using the overlay method as described [3]. Genes were scored as a hit if any abnormalities were observed in acini or lumen structure formation compared to vector alone. To assess lumen and acini formation, cells were stained with 0.5ng ml$^{-1}$ 4',6-diamidino-2-phenylindole (DAPI). Genes considered validated migration hits were also examined with 10A.B2 cells in three-dimensional basement membrane cultures with and without AP1510.

Transwell migration assay. 10A.B2 cells expressing genes from the BC1000 collection were starved overnight in assay media. $1 \times 10^5$ cells were added to the top chambers of 24-well transwell plates (BD, 8μm pore size), and assay media, with or without 500nM AP1510, was added to the bottom chambers. Genes were scored as a hit if there was an observed increase the number of migrating cells compared to vector alone.

3. RESULTS

To facilitate the application of functional genomics in breast cancer research, we have created the BC1000, a sequence-validated and individually arrayed cDNA collection of 1000 genes associated with breast cancer. In summary, the ability to capture human genes through PCR was a strong predictor of cloning success rate. Extensive testing and analysis of various conditions revealed, not surprisingly, that the most important PCR parameter was template. We observed a cumulative estimated mutation rate of 1 base change per 1000 nucleotides. Roughly 40% of observed base pair differences between the observed and expected clone sequence were previously described in the GenBank database, such as alternate splice forms and allelic variants.

Using a subset of genes in the BC1000 collection, we conducted three complex phenotypic assays with a nontransformed mammary epithelial cell line, MCF-10A, to investigate the genetic events involved in the different stages of breast carcinogenesis. When cultured in three-dimensional basement membrane gels, MCF-10A cells form growth arrested three-dimensional structures that resemble the mammary acini of the human adult breast. Subsequent activation of ErbB2 reinitiates cell proliferation and alters acinar and luminal structures. First, we conducted two screens to identify proteins that, when overexpressed, induced premalignant phenotypes similar to those observed with ErbB2 activation, specifically cellular proliferation and/or disruption of acinar morphogenesis. Second, we screened for genes that induced cell migration, independently or cooperatively with activated ErbB2, a phenotype characteristic of malignant
cancers. All three assays yielded well-studied genes expected to be involved in these processes, such as RAF1, JUNB, FOS, and GRB2, supporting the validity of our cell-based screening approach. Furthermore, we identified potentially novel functional activities for known breast cancer-associated genes. Among the three assays, the migration assay with activated ErbB2 yielded the largest percentage of hits, as well as most intriguing candidates to pursue.

**Table 1. Percentage of genes that scored as hits for each assay.**

<table>
<thead>
<tr>
<th>Assay</th>
<th>% Hit rate (no. hits/no. screened)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migration¹</td>
<td>17.3% (44/253)</td>
</tr>
<tr>
<td>Proliferation</td>
<td>4.6% (11/237)</td>
</tr>
<tr>
<td>Morphogenesis</td>
<td>11.7% (31/265)</td>
</tr>
</tbody>
</table>

¹ Validated migration hits with in 10A.B2 cells with activated ErbB2.

**4. CONCLUSIONS**

We have built an arrayed cDNA collection of breast cancer-associated genes, known as the BC1000 (www.hip.harvard.edu). To our knowledge, the BC1000 represents the first publicly available disease-based cDNA collection. Here, we have used this resource in three complex cell-based assays to investigate the genetic events involved in the different stages of breast carcinogenesis. Through this comprehensive approach, we identified potentially novel functional activities for known breast cancer-associated genes, as well as provided insight into the multiple biological effects of various candidate oncogenes. Subsequent analyses of the mechanisms underlying these findings may provide important insight into the process of breast cancer development.


CLINICALLY RELEVANT MEDICAL IMAGE COMPRESSION

MATTHEW J. ZUKOSKI, WILKES UNIVERSITY TERRANCE BOULT, U. COLORADO AT COLORADO SPRINGS TUNÇ IYRIBOZ, PENN STATE COLLEGE OF MEDICINE

ABSTRACT. As medical/biological imaging facilities move toward complete filmless imaging, compression will play a key role. Although lossy compression techniques yield high compression rates, the medical community has been reluctant to adopt these methods, largely for legal reasons, and have instead relied on lossless compression techniques which yield low compression rates. The true goal is to maximize compression while maintaining clinical relevance and balancing legal risk. It is important to figure out how to preserve the quality of the image at a level that is necessary for both what is clinically relevant and what doctor and lawyers feels is necessary for liability reasons, a criterion that may transcend what is visually perceptible or even actually necessary. This paper proposes a novel model-based compression technique that makes use of clinically relevant regions as defined by radiologists. Lossless compression is used in these clinically relevant regions with high quality and lossy compression is used everywhere else (i.e., the areas that are not clinically relevant). In our ongoing effort, the clinically relevant regions will be predefined on standard film types in an atlas. To demonstrate the potential savings, this paper considers simpler regions and shows the significant savings achieved.

1. INTRODUCTION

Medical image compression plays a key role as hospitals move towards filmless imaging and go completely digital. Image compression will allow Picture Archiving and Communication Systems (PACS) to reduce the file sizes on their storage requirements while maintaining relevant diagnostic information. Teleradiology sites benefit since reduced image file sizes yield reduced transmission times. Even as the capacity of storage media continues to increase, it is expected that the volume of uncompressed data produced by hospitals will exceed capacity and drive up costs.

This paper will propose an approach to improving the performance of medical image compression while satisfying both the medical team who needs to use it, and the legal team who needs to defend the hospital against any malpractice resulting from misdiagnosis due to faulty compression of medical images. The improved compression performance will be accomplished by making use of clinically relevant regions as defined by physicians. Images taken of patients will be aligned to pre-stored image models stored in an atlas. The atlas will contain models of typical classes of images. If we are trying to compress a chest x-ray image, then it will be matched with a pre-stored chest x-ray model that is stored in the atlas. If we are trying to compress an x-ray of the right hand, then it will be matched with a pre-stored right hand x-ray model that is stored in the atlas. The atlas will be discussed in more detail later in this paper. Once an image is aligned to its corresponding model in the atlas, the two can then be aligned and the clinically relevant regions defined on this atlas image will be used to define the relevant region on the newly scanned patient image. There are varied approaches to performing the alignment such as maximization of mutual information and deformable contour modelling.

This paper focuses on the potential gains to document the advantages of this approach. Lossless compression will be applied in the clinically relevant areas and lossy compression will be applied in the other areas. One of the reasons to use lossless compression in the relevant areas is not because radiologists think it is fine based on ROC curves. From our interviews, most doctors would prefer using lossy compression with the quality level on high. This would yield an image that is much smaller than it’s lossless-compressed counterpart and the image will be “visually lossless”. However, the lossy compressed image, when decompressed, is not identical to the original (i.e., there was loss) and the lawyers have a problem here. So this makes the problem not just what is perceptibly ”good enough” (i.e., what is ”visually acceptable”), but rather, what is ”clinically relevant.”

In Section 2, a very brief overview of medical image compression will be given. This includes lossy and lossless techniques. In Section 3, you will find an overview of the overall approach. This section will include a discussion of how the atlas will be created, and how the alignment of images to their corresponding models will be performed. An overview of alignment will be discussed here as well. The paper ends with preliminary results using simpler regions that document the potential savings.

2. MEDICAL IMAGE COMPRESSION

Before the various image compression techniques are discussed, consider the motivation behind using compression. A typical 12-bit medical x-ray may be 2048 pixels by 2560 pixels in dimension. This translates to a filesize of 10,485,760 bytes. A typical 16-bit mammogram image may be 4500 pixels by 4500 pixels in dimension for a filesize of 40,500,000 (40
megabytes)! This has consequences for disk storage and image transmission time. Even though disk storage has been increasing steadily, the volume of digital imagery produced by hospitals and their new filmless radiology departments has been increasing even faster. Even if there were in nite storage, there is still the problem of transmitting the images. Many hospitals have satellite centers or clinics in small towns and remote areas to make it convenient for patients that have a hard time travelling the distance to the hospital, especially for diagnostic procedures. These hospitals make use of teleradiology applications which allow the clinic staff to operate the clinic without the need for a radiologist to be present. Instead of a diagnostic radiologist, a technician or basic radiologist in the clinic can take the x-ray and send the image through a network connection to the hospital where the diagnostic radiologist can read the image and send back a diagnosis. But there is a problem, especially in emergency situations where time is of the essence, because a 10M image will take approximately half an hour using a high-speed modem. Broad-band connections such as T1 lines improve the situation, but many clinics are in such remote areas, that it is cost prohibitive to set up high speed lines, and the 56K modem is the most viable option. Cable Models or DSL typically have very asymmetric performance with high-speed downlinks and much slower uplinks, and the uplinks are what would limit the satellite facility. Even at standard Cable/DSL uplink speeds, a single 10 MB single radiograph will take over 5 mins to send. While that may sound reasonable, keep in mind that the patient is often asked to remain in the imaging apparatus until the radiologist has con med the data is suf cient. So compression is not just about the storage costs, it is also about transmission time, imaging apparatus utilization, and convenience/comfort of the patient. Compression techniques can reduce size and transmission time, thus improving overall care.

Image compression techniques take advantage of redundancy that occurs. There are different types of redundancy. Each compression methodology will exploit one of these redundancies. The different types of redundancies are, spatial, temporal and spectral. The research presented here will focus on the rst of these three types of redundancies although the techniques can be used in the others. It will make use of spatial redundancies since static spatial x-rays will be used. These are still the most dominant type of medical imaging data used today. In future work, the issue of using temporal redundancy will be explored as the proposed approach naturally applies on a dataset of images taken of one patient over time, and can be applied to video data as well.

Images can be compressed using lossy or lossless techniques. Lossless techniques allow the image to be compressed, then decompressed back to the original state of the image without any loss of data [5]. These methods are sometimes called reversible compression methods. Compression rates for lossless techniques vary but typically are around 2:1 to 3:1. On the other hand, lossy techniques do not allow for exact recovery of the original image once it has been compressed. These methods are sometimes called irreversible compression methods. But these techniques allow for compression rates that can exceed 100:1 depending on the compress quality level and the image content. At high quality lossy levels (10:1 to 20:1), compression rates much greater than those obtained by lossless methods can be obtained while achieving visually indistinguishable results. That is, the human eye cannot detect a difference between the original image and the compressed-then-decompressed image with the lossy compression method. However, the medical community has been very reluctant to adopt lossy algorithms in clinical practice. This is because of the legal questions raised and the regulatory policies set by agencies such as the Food and Drug Administration. To date, there is insuf cient clinical research on the use of lossy compression applied to medical images. The new compression approach, which will be proposed here utilizing a hybrid lossy/lossless method, can be made all lossy or all lossless.

The most popular compression algorithms in use today in the medical community are lossless JPEG (Joint Photographic Experts Group) [9] and lossless Wavelet. JPEG has been adopted by the Digital Imaging and Communications in Medicine (DICOM) group in their widely adopted DICOM image le format, but the wavelet compression algorithm is gaining ground. In fact, the DICOM group has announced that the next release of the DICOM format will utilize the JPEG 2000 standard, which uses wavelet compression.

2.1. Medical Compression Research. There has been numerous compression research studies examining the use of compression as applied to medical images. The papers can be categorized as focusing on just a lossless compression method, on just a lossy compression method, or focusing on both. Most have focused on lossless algorithms since the medical community has
been reluctant to adopt lossy techniques due to the legal and regulatory issues that are raised, but this situation may start to change as more lossy research is performed.

Lossless image compression is typically performed in two steps, decorrelation and coding. Image decorrelation attempts to reduce the redundancy within the image. There are several common approaches that have been taken in the literature to perform this redundancy reduction step including differential pulse code modulation, hierarchical interpolation, bit-plane encoding, and multiplicative autoregression. Several popular approaches for encoding are Huffman encoding, Lempel-Ziv encoding, arithmetic encoding, and run-length encoding.

Lempel-Ziv is used by Unix in the compress and gzip programs. It is also used in the GIF le format. The Huffman and Lempel-Ziv encoding methods were compared as applied to MRI images in [7]. It showed that Lempel-Ziv encoding methods achieve higher compression than compression ratios resulting from using Huffman encoding.

As mentioned earlier in this proposal, lossless methods are preferred in the medical community. Of these methods, JPEG and Wavelet are most popular. These two compression methods actually gained widespread acceptance as lossy methods. However, each can be made lossless which is the preferred style in medical imaging.

The transform based lossy methods involve three stages: transformation, quantization, and lossless coding. They divide the image into subimages (for example, 8 x 8 blocks in the DCT used in JPEG), and perform a transformation on the subimage, quantizing, and coding the resulting coefficients using some coding scheme such as Huffman, arithmetic, or run-length. Discrete cosine transforms are closely related to the Fourier transform and produce similar results. DCT’s convert data from the spatial domain into the frequency domain.

DeVore [3] showed that the wavelet transform is a promising tool for image compression providing high rates of compression while maintaining good image quality. In 1999, we performed a quantitative comparison of three lossy compression methods (one wavelet and two JPEG) as applied to a variety of 12-bit medical images in conjunction with the Department of Radiology at the Hershey Medical Center (Hershey, PA) [4]. This work shows the quality of JPEG and wavelet-based compression (which is what will be used in the remainder of the study).

With regard to clinically relevant region encoding, not much has been published. In 1994, [2] made use of regions of interest using subband analysis and synthesis or volumetric datasets using wavelets. They followed up this work with [1] by using structure preserving adaptive quantization methods as a means of improving quality for a compression rates in the regions of interest. But all of their effort was on lossy approaches.

In the most relevant work, [6] developed a region based coding approach. They discussed two approaches: one uses different compression methods in each region such as “contour-texture” coding and subband decomposition coding, and the other uses the same compression method in each region such as the discrete cosine transform but with varying compression quality in each region such as by using different quantization tables. They used two multiresolution coding schemes: wavelet zerotree coding and the S-tranform, and considered only 8 bit images. In their implementation, the regions of interest were selected by hand.

3. NEW APPROACH AND PRELIMINARY RESULTS

A new model-based approach to medical image compression through the use of image registration is proposed. An image which needs to be compressed, will be aligned to an image of its own type pre-stored in an atlas (such as the head or chest). Once an image is registered (i.e. aligned) two possibilities exist. The simpler approach is simply to read off the “relevant” regions and then use lossy compression in relevant regions and lossy compression in the others. The alternative is that the new image can be subtracted from the pre-stored atlas image generating a residual image. This residual image will be compressed (lossless in clinically relevant regions, lossy in the others). If the alignment is done well, the residual information is minimized, thus yielding higher compression.

The regions will be needed to classify areas of the image into those that are clinically relevant and those that are not clinically relevant. These regions are stored in the atlas and have been pre-stored by radiologists. The areas can be considered as initial or default. It is possible that the physician currently working with the image to be compressed may feel that these default regions are inappropriate and may want to modify them. The proposed system should allow for such interaction and provide means for the physician to override the default regions and define new ones.

Lossless compression will be used in the clinically relevant regions and lossy compression will be used
in areas that are not clinically relevant. Lossless and lossy compression were discussed in detail in the Introduction section. Most lossy compression algorithms, such as JPEG, utilize a compression amount parameter that defines the amount of compression, and hence degradation, used on the image. Since the clinically relevant regions, by definition, delineate the regions that the physicians care about, the amount of compression performed in the clinically non-relevant regions can be made very high. These regions will basically be used to provide a frame of reference to the physician and so long as the physician can make out in general what these clinically non-relevant regions are, then all is well.

A pre-stored database of common lm types such as x-rays of the head and chest can be stored. These will be used as templates and the collection will serve as the atlas. When an x-ray is taken, the technique will be told what type of x-ray was just taken. Next, the x-ray is aligned via the proposed deformable object model matching mutual information of the image against the pre-stored template. Now the two images are subtracted. Whatever is left is considered a residual. The residuals are encoded by compression. If the alignment has been done well, the residuals should be minimized which means the compression has been maximized.

This image alignment model is based on a hybrid registration technique that makes use of mutual information maximization between two images as an initial step, followed by another methodology based on deformable modelling.

Reconstruction is straightforward. The compressed residual image is rst decompressed. Then the decompressed residual is added back to the template according to the model and the original image should be obtained. This would be bene cial for teleradiology applications since only the compressed residual image needs to be stored along with the atlas.

One advantage of this method is that it does not matter near as much what the underlying data type is. It can be 2D or it can be 3D. Of interest is the level of compression data rate that can be achieved. It is possible with this technique to get very high compression rates with good alignment.

Another advantage of this method is that the residuals of residuals can be examined for possible improvement in storage savings. Consider a person whose x-rays have been taken over time. The rst x-ray can be stored after alignment. Then the difference between the second x-ray and the rst x-ray after alignment is computed. This is considered the residual of residuals. But there are two problems with this idea. The computing cost increases, and all of the residuals are useless if you lose the starting image. Long term x-ray differences can also be explored.

Whether complex contours will be useful is a research study in itself, not a solution to presume. One could use an approach like MPEG, but some model that deforms smoothly would be needed, not arbitrary patches going everywhere, but one might take a cubic spline and define a grid such that the only thing that moves on the grid when maximization of mutual information is performed, are the nodes of the spline and when one moves the nodes on the spline, everything else will smooth out. If one has compression in a given area of the image, the next block on the grid has to move smoothly as well from one image to the next.

The more complex issues of deformable model matching are still being studied and will be reported in later papers. The algorithm proposed here will also make use of maximization of mutual information, [8].

The rst level of mutual information maximization would be the static rotation, translation, and scaling, which is shown in preliminary results. More advanced techniques will rst determine the scaling parameters, then the nodes in the image will be allowed to start moving around in a subsequent pass. The global rotation, translation, and scaling parameters have to be encoded as well as the deltas for the grid. Then the residuals from subtraction can be obtained.

To provide supporting evidence for this ongoing study, the idea of clinically relevant regions has been applied to several medical images and the le sizes are compared between the image entirely compressed lossy, and the image that used the combined lossless/lossy approach. Each image presented is represented at 8 bits per pixel (bpp) only for this proposal. The nal paper will show results from 12-bit images as well. The 8-bit results are below.

3.1. Example 1. The rst image is a chest x-ray that was digitized by primary capture. That is, the x-ray was captured directly from the patient using a high-resolution 12-bit digital x-ray scanner. The image was then downsampled to 8 bits. Figure 1 shows the rst example image with three clinically relevant regions defined on it. These 3 areas partition the image into 7 areas as shown on the image indicated by regions 1 through 7. Regions 3 and 5 have been marked by a radiologist to be losslessly compressed. Region 4
has been marked by the radiologist to be lossy compressed, but at a higher level than the rest of the image. A JPEG level of 50 (of 100) was chosen for region 4. The clinically relevant region 3 has dimensions 836 x 1344 and its upper-left coordinate is located at position 164,296. The clinically relevant region 5 has dimensions 760 x 1344 and its upper-left coordinate is located at position 1288,296.

So, using the clinically relevant region approach, the compressed size is 382006 bytes. This gives a compression ratio of 11.221 : 1 or 0.713 bpp. The clinically relevant region area takes up 16.6% of the entire image area. The compression ratio has improved by more than a factor of 6 from 4.337 bpp for pure lossless.

3.2. Example 2. The next image is an x-ray of the skull that was digitized by ”secondary capture”. That is, the x-ray lm was digitized by using a high-resolution 12-bit scanner, then resampled to 8 bpp.

The dimensions of this lm are 1188 x 1528. The image is again automatically partitioned by having clinically relevant regions needed by registering to an atlas. Here is the image and the areas needed around it. Note that the clinically relevant region extends vertically to the bottom of the image. Therefore, the partitioning results in only 4 areas in this example. The clinically relevant region 1 has dimensions 996 x 1078 and its upper-left coordinate is located at position 142,452.

The original size of the uncompressed raw image is 5111808 bytes using 8 bits per pixel. With the entire image compressed using lossless JPEG-2000, the compressed size is 564185 bytes. This gives a compression ratio of 9.061 : 1 or 0.883 bpp.

Regions 3 and 5 are compressed using lossless JPEG. Regions 1, 2, 6, and 7 are compressed using lossy JPEG at a compression level of 10 (out of 100). Region 4 is compressed using lossy JPEG at a compression level of 50. The results are shown in table 1.

![Figure 1. Partitioned Chest x-ray (2496 x 2048, 8 bpp)](image1)

![Figure 2. Partitioned Skull x-ray (1188 x 1528, 8 bpp)](image2)

<table>
<thead>
<tr>
<th>Region</th>
<th>Raw Bytes</th>
<th>Type</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>738816</td>
<td>lossy</td>
<td>3331</td>
</tr>
<tr>
<td>2</td>
<td>220416</td>
<td>lossy</td>
<td>740</td>
</tr>
<tr>
<td>3</td>
<td>1123584</td>
<td>lossless</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>387072</td>
<td>lossy</td>
<td>3993</td>
</tr>
<tr>
<td>5</td>
<td>1021440</td>
<td>lossless</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>602112</td>
<td>lossy</td>
<td>1478</td>
</tr>
<tr>
<td>7</td>
<td>1018368</td>
<td>lossy</td>
<td>2315</td>
</tr>
<tr>
<td>TOTAL</td>
<td>5111808</td>
<td></td>
<td>284259</td>
</tr>
</tbody>
</table>

Table 1. Region Compressions. Overall compression rate is .445 bpp, almost 2 times smaller than pure lossless.
So, using the clinically relevant region approach, the compressed size is 519371 bytes. This gives a compression ratio of 3.495 : 1 or 2.289 bpp. The clinically relevant region area takes up 59.0% of the entire image area. So even though the clinically relevant region is so large, the compression ratio has still been improved over the 2.358 : 1 ratio given earlier in this example.

3.3. Residual Image. The previous subsections provide overview of compression with respect to the clinically relevant regions on the raw image. An important question is if using the atlas we can ultimately produce better compression using a residual image. Then this residual image will be compressed in a similar region-based approach. For example Figure 5 shows the partitioned residual image produced when the image used in Example 1 is subtracted from the image obtained (see Figure 4) after matching the corresponding atlas image (see Figure 3) to the Example 1 image. The partitioning is the same as in Example 1. As before, the three clinically relevant areas partition the image into 7 areas as shown on the image indicated by regions 1 through 7. The clinically relevant region 3 has dimensions 836 x 1344 and its upper-left coordinate is located at position 164,296. The clinically relevant region 5 has dimensions 760 x 1344 and its upper-left coordinate is located at position 1288,296.

The original size of the uncompressed raw image is 5111808 bytes using 8 bits per pixel (bpp). With the entire image compressed using lossless JPEG-2000, the compressed size is 2987890 bytes. This gives a compression ratio of 1.711 : 1 or 4.676 bpp. Regions 3 and 5 will be compressed using lossless JPEG. Regions 1, 2, 6, and 7 will be compressed using lossy JPEG at level 10. Region 4 will be compressed using lossy JPEG at level 50. Here are the results.

So, using the clinically relevant region approach on the residual, the compressed size is 1380229 bytes. This gives a compression ratio of 3.704 : 1 or 2.160 bpp. The clinically relevant region areas takes up 42.0% of the entire image area. The compression ratio has
improved over 100% over the 1.711 : 1 ratio for the lossless compression of the residual image.

While only a single example, it is important to note that compared to applying the clinically relevant region approach to the original image, the use of the residuals may not be more effective. In this example, residuals produced a final result of 2.160 bpp compared to the 0.445 bpp achieved using the same clinically relevant regions on the original images. There are fundamental reasons to expect that the residual may not help. The first problem is that the residual approach must encode the atlas used and “transform” parameters. The more difficult problem is that even minor misalignments result in high-amplitude high-frequency data (i.e. the residual image looks mostly like edges), which are then harder to compress. This suggests that the simple 6-parameters of the transforms used for alignment are insufficient, and motivates the development of more general deformable models.

While the residual approach is still experimental, this paper has shown that the overall approach of clinically relevant regions has clearly demonstrated advantages over both traditional lossless compression and simple lossy compression, and that the residual approach holds some promise. A larger scale study in conjunction with Penn State University’s Hershey Medical Center is ongoing (and should be complete by the time of the symposium).

REFERENCES


E-mail address: zukoski@wilkes.edu & tboult@cs.uccs.edu & tunc@psu.edu
LP³ - The LEA Protein Prediction Pipeline
Martin Gollery, Barrett Abel, John Cushman
Department of Biochemistry, University of Nevada, Reno, NV 89557-0014, USA

Abstract
Late Embryogenesis Abundant (LEA) proteins are a particular class of hydrophilic proteins or hydrophilins, which are increasingly synthesized during desiccation of maturing seeds. LEA proteins are also found in abundance in vegetative tissues during dehydration, low temperatures, osmotic stress or ABA application (1,2). LEA proteins and other hydrophilins are ubiquitous in plants, and some classes of hydrophilins are also present in nematodes, bacteria and yeast. Despite the breadth of distribution and the conservation of their sequence structure, the biological role of most hydrophilins remains unknown. Gene expression studies show that hydrophilins respond to cold, salinity, osmotic stress or abscisic acid treatment, but our goal here was to produce a computational pipeline to predict the presence of hydrophilins in general, and LEA proteins in particular, in any given proteome without expensive laboratory preparations. Rather than rely on a single prediction tool to detect putative hydrophilins, we utilized a number of varied algorithms to provide a more reliable prediction. The ultimate goal of this project is to determine LEA proteins in a wide range of proteomes in order to provide a resource to others that will ultimately lead to the full elucidation of the role or roles of LEA proteins.

Introduction:
Several types of Late Embryogenesis Abundant proteins or hydrophilins have been identified in the public databases, including entries in the PFAM database, representing proteins primarily from plants, but including bacteria, yeast and algae. LEA proteins are most frequently synthesized at the onset of desiccation in maturing seeds and during times of water deficit in vegetative tissues. Many LEA proteins are regarded as novel hydrophilins both by their high hydrophilicity and high Glycine content (>8% of total residues). LEA proteins are thought to bind water molecules to minimize water loss, but several other roles have been proposed, including the ability to act as chaperones for other proteins, membrane stabilization, or ion sequestration. Specific expression of LEA proteins can be induced by water deficit, cold, high salinity, and osmotic stress. Despite the correlation of these varied stress conditions and LEA protein or hydrophilin expression, the exact functional roles of these proteins remains unclear.

Our goal with this project is to develop a tool that combines several existing methods with some novel algorithms to provide a useful tool for predicting LEA proteins across a
proteome. While our primary target organism is Arabidopsis, the LP3 may be used for any organism.

Methods

InterProScan:

The InterPro database (http://www.ebi.ac.uk/interpro/) currently recognizes four categories of LEA proteins: LEA protein 1 (IPR004864), LEA protein 2 (IPR004926), and LEA protein 3 (IPR005513), and LEA protein (IPR004238). Proteins may be compared against this dataset by downloading and installing the InterProScan program suite for local use, or searches may be run online at the InterPro website. The major advantage to running local searches is the ability to analyze large numbers of sequences in one large batch, although the computational complexity of the algorithms involved will impose a practical limit on the number of searches run.

HydroBLAST:

The Blastall program from NCBI was modified slightly to evaluate the Arabidopsis proteins in terms of hydrophilicity. This program, called HydroBLAST uses a modified scoring matrix, called Hydro, that utilizes the hydropathy of each residue based on the Kyle-Doolittle hydropathy values(5), normalized to all positive values. These values then ranged from 0-9, as opposed to -4.5 to 4.5 on the Kyle-Doolittle scale (see Figure 1).

Each mismatch was then assigned to a negative number in the scoring matrix. This ensured that the best hit from each query was itself in the target, and that the lowest score would be assigned to the most hydrophilic protein. The HydroBLAST output was then compared to that of the Hydrophil Perl script.

| A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  | *
| A | 6 | -1 | -2 | -2 | -3 | -1 | -1 | 0 | 2 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 0 | 0 | 2 | -3 | -1 | -1 | 0 | 2 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| N | -2 | 0 | 1 | -1 | -3 | 0 | 0 | 0 | -1 | -3 | -3 | 0 | -2 | -3 | -2 | -1 | 0 | -4 | -2 | 3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | -1 | 1 | -3 | 0 | 0 | 0 | -2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | 1 | 0 | -1 | -4 | -3 | 3 | -9 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 7 | -3 | -4 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -2 | -1 | -2 | -1 | 3 | -3 | -2 | -4 |
| Q | -1 | -1 | 0 | 0 | -3 | 1 | -2 | -2 | 0 | -3 | -2 | -1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | -3 | -1 | -4 |
| E | -1 | 0 | 0 | -2 | -4 | -2 | 1 | -2 | 0 | -3 | -2 | -1 | 0 | -1 | -3 | -2 | -2 | -1 | -4 | -1 | -4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 2 | -4 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -3 | 3 | -1 | -2 | -1 | -4 |
| H | 2 | 0 | -1 | -1 | -3 | 0 | 0 | 0 | -2 | 1 | 3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 3 | 0 | 0 | -1 | -4 |
| I | 1 | -3 | -3 | -3 | 1 | -3 | -3 | 4 | 9 | 2 | -3 | -1 | 0 | -3 | -2 | -1 | -3 | -1 | -3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | -2 | 8 | -2 | -2 | 0 | -3 | -2 | -1 | -2 | -1 | -1 | 4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -3 | -2 | -1 | 0 | -1 | -3 | -2 | 0 | -1 | -1 | -4 |
| M | -1 | 1 | 0 | -2 | 3 | -3 | -2 | -1 | -2 | -1 | 6 | 0 | 2 | -1 | -1 | -1 | -1 | -1 | 3 | -3 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | 3 | 1 | 0 | 0 | 0 | 0 | -7 | -4 | -2 | -2 | -1 | -3 | -1 | 3 | -3 | -1 | -4 |
| P | -1 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | 3 | -1 | -2 | 4 | 3 | -1 | -1 | -4 | -3 | -2 | -2 | 1 | -1 | -2 | -4 |
| S | -1 | -1 | -1 | 0 | 1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | -1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -2 | -1 | -4 | -2 | 0 | -1 | -1 | 0 | -4 | 4 | 8 | 0 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -4 | -2 | -3 | -2 | 1 | -1 | -2 | -3 | -4 | 2 | -3 | -4 | -3 | -2 | -4 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -3 | -2 | -1 | -2 | -3 | -2 | 1 | -1 | -2 | 3 | -3 | -2 | -2 | -3 | 1 | -3 | -3 | -2 | -1 | -4 | -4 |
| V | 0 | -3 | 3 | -3 | 1 | -2 | -3 | -3 | 1 | 2 | -1 | -1 | -2 | -2 | 0 | -3 | 1 | 9 | -3 | -2 | -1 | -4 | -1 |
| B | -2 | -1 | 3 | -4 | -3 | 0 | -1 | -1 | 0 | -3 | 4 | 0 | -3 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | -1 | -1 | -4 |
| Z | 0 | -1 | 0 | 1 | 3 | 3 | 4 | 2 | 0 | -3 | 3 | 1 | -1 | -3 | 1 | 0 | -1 | -3 | -2 | 2 | 1 | -4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

Figure 1: The HydroBLAST matrix.
# Matrix made by M. Gollery for hydrophyt
# Using modified Kyle-Doolittle statistics
# * column uses minimum score
# Scoring Matrix in 1/2 Bit Units

**Hydrophil Perl Script:**

To identify hydrophilic Arabidopsis proteins with a high Gly content, an ad-hoc Perl program written by Alejandro Garciarrubio (1) was used on all Arabidopsis proteins derived from the TIGR ATH1.pep database. For each sequence, the program records the AA composition and performs a hydrophilicity analysis based on the Kyte and Doolittle algorithm (5). The program scores each protein according to the selection criteria wherein each protein had an average hydrophilicity > 1 and > 8% Gly. The score for such proteins given by the following formula (100 x (%Gly - 8) x (average hydrophilicity - 1) was chosen to give zero at the threshold values and to grow larger as a protein becomes more distinguishable from the bulk of the non-hydrophilin proteins.

**POPPS:**

While the Smith-Waterman algorithm and its approximations such as BLAST and FASTA are useful for structured and closely related proteins, the LEA protein or hydrophilins are apparently too diverse and unstructured to be characterized in this manner. Unstructured proteins are able to undergo considerable sequence rearrangement without loss of function, and this rearrangement makes it difficult for sequence similarity programs to identify homology. Therefore, alternative approaches must be used in an attempt to identify related LEA or hydrophilin family members. One such alternative approach is the Protein or Oligonucleotide Probability Profile (POPps) method (4), which does not attempt to construct an alignment at all. Rather, the POPps system takes a count of the number of mono- di- and tri- peptides in a sequence and compares this to an ‘average’ count of each of these peptides and records those that are under- or over represented. The ‘profile’ then is the list of peptides that are divergent from the average in the database. It is this profile, not the sequence itself, which is used to impute similarity. POPps promises to be an important new method for clustering as well as for non-linear sequence analysis.

**HydroHammer:**

Multiple Sequence Alignments of the Arabidopsis hydrophilins showed that they fell naturally into 11 groups, or families. These groups were then built into Hidden Markov Models with the HMMer package from Sean Eddy at Washington University (http://hmmer.wustl.edu/) and calibrated using hmmscan. These Models have been built into a small HMM database called HydroHammer, and is available upon request. This database is available to run searches online (http://chloroplast.biochem.unr.edu) for those who have accounts at the Nevada Center For Bioinformatics. This server is an
accelerated DeCypher system from TimeLogic, to provide extremely high throughput HMM searches.

Hidden Markov Models are an efficient method for comparing data against protein families, with an excellent blend of selectivity and specificity. The models in HydroHammer will be improved iteratively. As more members of each family are discovered from different organisms, they will be added to the statistical model and that model will then be recalibrated to include that new data.

**Results**

The Hydrophil perl script provided slightly different values from the HydroBlast matrix, due to differences in the scoring methods, but we found that the order was virtually identical when the proteins were sorted from greatest hydrophilicity to least hydrophilic. We appreciated several of the additional features found in GarciaRubio’s script, such as the percent Glycine content, and so recommend that hydrophil.pl be used rather than HydroBlast for general use. The perl script also proved to be considerably faster than BLAST in our tests. We may add this scoring matrix to our TeraBLAST server in the future, however, which would enable us to quickly check large proteomic datasets.

POPPs promises to be an extremely useful algorithm, providing a truly novel approach to the analysis of sequence data. We have encountered many practical problems with scalability, however, and the program frequently hangs with larger datasets. Changing to stricter thresholds provides some relief to this problem.

We have identified 90 hydrophilins in the Arabidopsis proteome, which is a considerable increase over the 32 that are currently annotated. These have been grouped into 11 families based on sequence similarity. We expect to find similar if not greater results with other, less well-annotated datasets.

We have begun molecular modeling of the group 1 LEA protein family, and have begun to uncover some interesting characteristics of these molecules. This work is in the early stages, and we expect that this will form a second-stage of the overall goal of elucidating the function of the LEA proteins.

The purpose of the LP\(^3\) is the discovery and analysis of LEA proteins. After the identification and classification of the proteins is complete, a systematic functional analysis program will be initiated involving gain-of-function analysis in yeast and loss-of-function analysis in Arabidopsis using traditional knock-outs or RNAi approaches.
Table 1: The LEA Protein Prediction Pipeline

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>InterProScan</td>
<td>Pipeline</td>
<td>Publicly available, multiple methods, widely used, online or local implementation.</td>
</tr>
<tr>
<td>HydroBlast</td>
<td>Heuristic</td>
<td>Easily installed, yet slower and less information than Hydrophil.</td>
</tr>
<tr>
<td>Hydrophil</td>
<td>Content analysis</td>
<td>Perl script, runs anywhere, provides much useful information.</td>
</tr>
<tr>
<td>POPPs</td>
<td>Profile</td>
<td>Analyzes a probability profile of Oligonucleotides.</td>
</tr>
<tr>
<td>HydroHmmer</td>
<td>Dynamic</td>
<td>Statistical representation of protein families, used with HMMPfam program.</td>
</tr>
</tbody>
</table>

Acknowledgements

Martin Gollery and the Nevada Center For Bioinformatics is supported by Award P20 RR16464 from the National Center for Research Infrastructure through the Biomedical Research Infrastructure (BRIN) program.

References


3) EBI Proteome Analysis Database [http://www.ebi.ac.uk/proteome].


F.Servant, C.J.A.Sigrist, E.M.Zdobnov),
"The InterPro database, an integrated documentation resource for protein families,
domains and functional sites",

7) Sonnhammer EL, Eddy SR, and Durbin R (1997) Pfam: a comprehensive database of
protein families based on seed alignments. *Proteins* 28:405-420
Jumpstarting Phylogenetic Analysis

Mark Clement, Quinn Snell, Kevin Seppi, Keith Crandall
3360 TMCB
Brigham Young University
Provo, Utah 84602
clement@cs.byu.edu

Abstract
When a new epidemic strikes, it is often important to determine the relationship between the current organism and others that have been successfully treated previously. The phylogenetic analysis problem generates the most likely family tree for a group of organisms based on DNA sequence data. This process can take a prohibitively long period of time with current algorithms. If trees resulting from prior searches are used to seed the search, correct trees can be found much more quickly. This jumpstarting algorithm can generate superior phylogenetic solutions much more quickly than existing algorithms.

1. Introduction

Phylogenetic analysis has become an integral part of many biological research programs. These include such diverse areas as human epidemiology [20,21], viral transmission [22], biogeography [23], and systematics [24]. With the advent of high speed sequencing equipment, an increasingly large volume of sequence data is becoming available. Scientists should be able to take advantage of this data and also of the research that others have performed. For example, when a new virus is detected, it should be possible to build a phylogenetic tree (an evolutionary history) containing all related viruses and the unknown variety in order to answer questions such as:

- Where did this virus come from?
- When did this virus arrive in the human population?
- What are the related species from which we might derive ideas about appropriate antibodies for testing and remedies for treatment?
- Has this virus been genetically modified through natural or human induced recombinant technology?
- How is this virus evolving and what genetic changes occurred to allow it to successfully enter the human population? This allows us to gain insights into how we can prevent future outbreaks.

Unfortunately, this kind of phylogenetic search is currently computationally infeasible. The time it takes to perform a complete search using maximum likelihood exceeds several months with even a small number of taxa. In the case of the SARS epidemic, and others like it, treatment information must be available in days or at most weeks in order for appropriate action to be taken. Much of the problem comes from the culture and software design for most phylogenetic software packages [14,18,27]. These packages require the user to start a search from scratch every time a new sequence is added to the search (this is exactly the situation when a new antigen is observed). The software packages also do not allow users to share partial trees that could speed up the phylogenetic search process. This creates a culture where investigators see little or no benefit to collaboration in phylogenetic research.

1.1 Phylogenetics

The branching pattern of ancestor/descendant relationships among species or their parts (e.g., genes) is a phylogeny. Researchers attempt to estimate these historical relationships by examining character evolution
using a tree — a mathematical structure used to model the actual evolutionary history of species or their parts [28]. These inferred trees (historical branching relationships) can be represented as cladograms, where branch lengths are arbitrary and only the branching order is significant, or as phylogenies, where the branch lengths are proportional to the amount of evolutionary change along the branch.

Phylogenies were historically used to classify organisms into natural evolutionary groups based on these ancestor/descendant relationships. Indeed, great effort is currently being spent on estimating the “tree of life” to quantify the biodiversity of our planet. However, phylogenies have also spread in use as the utility of the evolutionary framework for numerous other disciplines becomes increasingly obvious. For example, phylogenies are now being extensively used in the biomedical sciences including developmental biology, genomic biology, infectious disease, virology, and human genetics.

Phylogenies have become essential tools in the study of the molecular epidemiology of disease agents [30]. A prime example of the troubles encountered when the phylogenetic approach is ignored comes from the outbreak of the West Nile Virus in New York City. This virus was responsible for multiple deaths in New York, yet the Centers for Disease Control and Prevention (CDC) initially misdiagnosed the causative agent as St. Louis encephalitis due to their lack of an appropriate phylogenetic comparison [25]. The study of origins, spread, and diversity of pathogens are clearly evolutionary questions. Only after the serological evidence was coupled with strong phylogenetic evidence was the etiological agent responsible for the encephalitis outbreak in New York correctly identified as the West Nile Virus [29].

Phylogenetic estimation is accomplished by optimizing character change relative to some criterion over a tree. The tree for which the character data show the best optimization is the preferred tree. Two of the principle optimization criteria used by researchers are maximum parsimony and maximum likelihood. The parsimony criterion attempts to minimize the number of changes among a tree for shared-derived characters, while likelihood attempts to maximize the probability of change for all characters relative to some model of evolution. Each criterion has its own strengths and weaknesses. For example, maximum parsimony can incorporate insertion/deletion (indel) events and have asymmetric changes (e.g., a change from character A to character B is not the same as a change from character B to character A), whereas current implementations of maximum likelihood cannot accommodate these biological realities. Likewise, maximum likelihood can account for heterogeneity in evolutionary rates and multiple changes at the same character position, whereas maximum parsimony cannot. Thus there is, often times heated, discussion about appropriate methods to use to estimate phylogenetic relationships.

Another reason there is such debate about phylogenetic methods is that their performance varies depending upon the type of data used, the number of sequences involved, and the depth of the evolutionary relationships to be inferred. Exact searches, those that explore every possible tree topology for a given optimality criterion, are only possible for a very small number of taxa (on the order of 20-30). This limited search is due to the rapidly increasing number of possible trees with a modest increase of taxa [26]. The total number of (unrooted, strictly bifurcating) trees for $T$ taxa is

$$B(T) = \prod_{i=3}^{T} (2i - 5).$$

So, for example, with only 50 sequences, there are $3 \times 10^{74}$ possible trees. For the tree of life, there are estimated to be well over 10 million species, yet for 10 million sequences there are $5 \times 10^{68.667.340}$ possible trees! Therefore, the phylogeny problem is a particularly tough one that is well suited for distributed technology (because one performs the same calculations over different, independent, tree topologies) such as web based systems that utilize distributed resources.
Phylogenetics has become an active field in and of itself [31]. It is an extremely exciting field where talents in mathematics, computer science, and biology can be brought together to work on the problem of inferring historical relationships. A survey of the recent literature in many fields in the biomedical literature will attest to the ever increasing applicability of phylogenetic analyses to these fields.

2. Jumpstarting

The jumpstart algorithm first searches a database for similar searches that have been performed previously. The best tree from the prior search with the maximum subset of common taxa is then used as a starting point for the next search.

**Jumpstarting Algorithm:**

1) Let \( T = \{ x \mid x \text{ is in the set of taxa involved in the new search}\} \).
2) Query the data base for prior searches with the set of taxa \( S_i \) where at least one of the taxa in the prior search is the same as the new search \( \exists x (x \in T \text{ and } x \in S_i) \).
3) For each of these prior searches on taxa \( S_i \), determine the intersection \( I_i = T \cap S_i = \{ x \mid x \in T \text{ and } x \in S_i \} \).
4) Find the intersection \( |I_i| \) with the largest number of common taxa.
5) Use the Newick parenthetical notation for the best tree from this maximal intersection as the base tree for the new search.
6) Add taxa \( x \in T \text{ where } x \not\in S_i \) in positions where they minimize the length of the tree.
7) Begin a normal search with the tree from the previous search as a starting place.

Figure 1.1 provides a concrete example of the jumpstart algorithm. In this example, User A on Peer 1 has performed a search resulting in Tree X version 1.1. The following steps are included in the algorithm.

1) User B on Peer 2 prepares a set of taxa that will be used in a phylogenetic search and creates the data structure for Tree Y, Version 1.1. A query is sent to peer machines to determine if searches have already been performed with some of these taxa.
2) Peer 1 has Tree X Version 1.1 which matches the criteria in the query. This tree is returned to Peer 2.
3) Peer 2 uses Tree X Version 1.1 combined with other local taxa to jumpstart a phylogenetic search.
4) After expending significant computational resources, User B generates Tree Y Version 1.2 which refines the relationships between taxa in Tree X as well as Tree Y. This version of the tree is entered into the database.
5) User A has a reference to Tree Y since a subtree of Tree X was used as a jumpstart point for Tree Y. When Tree Y Version 1.2 is generated Peer 1 can send a query for derivative trees of Tree X Version 1.1.
6) Peer 2 will return Tree Y Version 1.2. User 2 may decide that all of the relationships contained in Tree Y Version 1.2 should not be made public. In this case, the subtree containing only the nodes originally found in Tree X would be returned.
7) Peer 1 receives the refined relationships and can create Tree X version 1.2. This tree can be used for future searches.

In this example interaction, both User A and User B have benefited from the collaboration. The tree returned from Peer 2 can be used, or discarded depending on the value that User A places on the results. User B has been able to cut months off of his search time because of the initial jumpstart tree he/she was able to derive from Tree X Version 1.1.
Figure 1.1 Example interactions between jumpstarting peers
3. Experimental Results

Typical phylogenetic research involves the collection of data, sequence alignment, and phylogenetic analysis. Regrettably, the results obtained in the last two stages in this process are not generally able to be used in subsequent analyses. For example, when a new sequence is added, a new alignment must be computed, and the old trees that were generated in phylogenetic analysis cannot currently be used as a starting point. Both alignment and phylogenetic analysis are time-consuming processes (some analyses have been known to take months). When a researcher adds a sequence to the analysis, all the previous time spent in computation is essentially thrown away because the researcher must start over. This drastically slows the scientific process. Our research group has been studying ways to eliminate the wasted time and take advantage of previous analyses.

The most promising way to “jumpstart” phylogenetic analysis is to use previously computed trees. Basically, the idea is to start where you or another researcher left off rather than starting over. Typically, researchers a) don’t save previously computed trees, and b) don’t have easy means whereby they may use previously computed trees. To determine the benefit of jumpstarting, we randomly selected a sequence to remove from the target dataset. Phylogenetic analysis was then performed on the remaining data and the best trees were obtained. Finally, the removed sequence was added back into the dataset and new trees were generated using our software in conjunction with PAUP*[14]. The time to generate the most optimal tree using jumpstarting was compared with the time required to start from scratch (see Table 3.1). More optimal trees have shorter length values.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Strategy</th>
<th>Paup* time (seconds)</th>
<th>Paup* length</th>
<th>Jumpstart time (seconds)</th>
<th>Jumpstart length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zilla</td>
<td>Parsimony</td>
<td>288054</td>
<td>16220</td>
<td>510</td>
<td>16218</td>
</tr>
<tr>
<td>HIV</td>
<td>Parsimony</td>
<td>2410</td>
<td>110709</td>
<td>1816</td>
<td>110658</td>
</tr>
<tr>
<td>SARS</td>
<td>Likelihood</td>
<td>30605</td>
<td>542771.9</td>
<td>127</td>
<td>542771.9</td>
</tr>
</tbody>
</table>

Clearly, jumpstarting is advantageous. In all cases, jumpstarting resulted in trees that could not be found by starting from scratch. In the case of the Likelihood search with SARS data, the collaborative search was performed 240 times faster than the search without jumpstart trees. In the experiments with parsimony, not only was the search time drastically reduced, but the quality of the trees (smaller tree lengths) was also improved. More research is needed with a larger number of data sets to determine average performance improvements, but initial results indicate that collaboration in this area can have a significant impact on the quality of and time taken for phylogenetic searches.

Jumpstarting is advantageous when adding sequences to an existing analysis. However, this is not the only use case. More typically, a researcher may request data from the database and wish to begin computation by taking advantage of these data. A user can request all trees that contain certain taxa or sequences from the database. However, these trees may also contain extraneous taxa. One approach is to simply strip out the extraneous samples and use the resulting trees as a starting point. A consensus tree could also be used.
Figure 3.1 Comparison of various jumpstart methodologies

Figure 3.1 demonstrates the different jumpstarting possibilities available. A researcher may choose to start computation based on one of the most optimal (Smallest) trees returned from the query. Optionally, a consensus tree may be created and used for jumpstarting the new search (Strict, 50%). Preliminary studies show that creating a majority rule consensus tree (50%) from the collection of most optimal trees returned by the jumpstart system seems to be the best option. Figure 3.1 shows that this method found the most optimal tree in 17 minutes, whereas the other choices took at least 57 hours. Increases in performance such as this are vital to advances and to apply the power of the phylogenetic approaches to studies in biomedical research.

4. Conclusions
Jumpstarting is an effective algorithm for improving search times and tree quality in phylogenetic analysis. It can be combined with various search algorithms to deal with important medical problems. Future work will investigate ways of mining the database and combining existing trees to provide the best starting point for future searches.
5. References

5. NIH Collaboratory Project, http://collaboratory.psc.edu
SourceForge (2003), http://sourceforge.net/
BugZilla (2003), http://bugzilla.mozilla.org/
Recognizing TATA Promoters Based on Discriminating Frequency Analysis of Neighborhood Tuples

Raja Loganantharaj  
Center for Advanced Computer Studies  
University of Louisiana at Lafayette  
logan@cacs.ull.edu

M E Karim  
Center for Advanced Computer Studies  
University of Louisiana at Lafayette  
mek@cacs.ull.edu

Arun Lakhotia  
Center for Advanced Computer Studies  
University of Louisiana at Lafayette  
arun@cacs.ull.edu

Abstract  
Using plant database we investigate the distribution of different tuples of small length for TATA and TATA-less promoters. Results show that in the neighborhood of TATA patterns some of the tuples demonstrate discriminating distribution for TATA and TATA-less promoters. Based on this observation we perform probabilistic learning to recognize TATA promoters and are able to reduce the false positives significantly keeping the true positives comparable to existing approaches.

Keywords  
Promoter, TATA-box, Bayesian learning.

1. Introduction

Prediction of promoters is one of the core problems in bioinformatics. If we know the promoters then we can infer the location of the transcription start sites (TSS) and thus find the coded region in DNA. A TSS is usually associated with a Promoter Element (PE) like TATA-Box or Inr. Promoter prediction schemes attempt to identify these PEs to predict promoters.

Most of the existing methods for predicting promoters have a high false prediction rate. Many nucleotide subsequences that look like PEs are not really PEs. It is not completely understood why the same pattern is active as PE in some cases and is not in others.

Little is known about whether the presence or the absence of some neighborhood tuples have any effect in determining PEs. We have focused on studying a particular PE called TATA-box which is defined as 5’ – TATAWAW-3’ where W is either ‘A’ or ‘T’. We are applying statistical machine learning technique particularly the Bayesian learning algorithm for recognizing whether a core promoter with TATA active or not.

We used PlantProm DB (http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom), an annotated non-redundant collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site(s) (TSS) from various plant species, for present study. It was developed by the Department of Computer Science at Royal Holloway, University of London in collaboration with Softberry Inc. The current release of PlantProm DB contains 305 entries, of which 71 are monocot, 220 are dicot and 14 are other plants. We noticed that tuples of nucleotides of a small length have similar distribution patterns for both TATA and TATA-less promoters over an extended region around TSS, however, discriminating distributions at the neighborhood of TATA. We utilize this information to build statistical tools for recognizing TATA-boxes.

This paper is organized as the following. We provide a brief overview of the related work in Section 2. This is followed by our approach and associated results in Section 3. Finally we conclude in Section 4.

2. Related Study and Existing Tools

2.1. Pattern Analysis

Zhang[1] studied 177 human non-redundant promoter sequences extracted from EPD48. He showed, using position-specific k-tuple feature variables, that some tuples appear with higher frequency over others. He also noticed that based on that observation a quadratic discriminant analysis method could be effective in identifying human core promoters.

Bajic et. al.[2] studied extended human core promoter regions covering [-70,+60] segment relative to the transcription start site of human promoters contained in EPD. They found that the most common transcription factor binding site appeared to be initiator and other less obvious sites were Spz1, E2F-1, ZF5, and C/EBP. The ‘cap' site was also in this most common group. The TATA-box was found in only 11.63% of all examined promoters. Additional analyses revealed that the other common promoter elements include AP-2, CdxA, Pax-2, SRY, STAT1 and STAT5A. It was also observed that a number of promoter elements show strong preference either for the GC-rich or the GC-poor core promoters.
2.1 Existing Tools

Fickett and Hatziioygou[3] made a comprehensive review of publicly available software tools for locating promoters in DNA sequence. Summary of their work is presented below:

**Audic/Claverie**

Audic and Claverie (1997) construct Markov models of vertebrate promoter sequences (based on EPD) and nonpromoter sequences. For an arbitrary test window a Bayesian choice is made between the promoter and nonpromoter hypotheses.

**Autogene**

Autogene includes a module for promoter recognition (Kondrakhin et al 1995). The program utilizes a set of 136 consensus sequences for transcription factor binding sites collected by Faisst and Meyer (1992). A training set of 472 promoters was taken from the EMBL Database, based on annotation in EPD and EMBL. The occurrence frequencies for each of the consensus sequences in 50 fixed length sub regions of the promoters was determined. In a test sequence, an occurrence of one of the consensus sequences in one of the sub regions was weighted according to the frequency with which it occurred in that sub region in a certain subset of the training set (determined by a clustering algorithm based on the consensus site occurrences) and the expected frequency of occurrence in random DNA.

**GeneID/Promoter1.0**

An unpublished promoter-finding algorithm, developed by S. Knudsen (Technical University of Denmark), is included in the GeneID e-mail server. According to the on-line documentation, “Promoters are predicted by a program called promoter1.0. It has been developed as an evolution of simulated transcription factors that interact with sequences in promoter regions.”

**NNPP**

NNPP (M. Reese) combines recognition of the TATA box and the Inr, using the time delay neural net architecture, which allows for variable spacing between the features.

**PromFind**

PromFind (Hutchinson 1996) is not based on any collection of putative transcription factor binding sites but, rather, on the differences in nucleotide hexamer frequencies (following Claverie and Bougueleret 1986) between promoters, protein coding regions, and noncoding regions downstream of the first coding exon. Training and testing sets were taken from some of the GenBank sequences with corresponding entries in EPD.

Among all sites in an input sequence where the promoter versus coding region discriminant exceeds a certain threshold, the site where the promoter versus noncoding region discriminant reaches its maximum (over the input sequence) is taken as a promoter.

**PromoterScan**

PromoterScan (Prestridge 1995) recognizes primate promoters by means of (1) the TATA PWM from Bucher (1990), and (2) the density of specific transcription factor binding sites. In calibration, occurrences of each transcription factor binding site listed in TFD was counted in EPD primate sequences and in primate nonpromoter sequences from Gen-Bank. The ratio of the densities of occurrence in each of these two sets is used as a weighting factor for that site. Then in application, the weighting factors for those sites occurring in the test sequence are combined with a TATA box score. The algorithm sometimes suggests a TSS and sometimes only gives a 250-bp window within which a core promoter sequence is thought to occur.

**TATA**

Bucher found that most TATA boxes were centered at a point 20–36 bp upstream of the TSS, so 28 bp downstream point of the center of the putative TATA box was taken as the predicted TSS.

**TSSG and TSSW**

TSSG and TSSW (Solovyev and Salamov 1997) both use the same underlying algorithm, which uses a linear discriminant function combining (1) a TATA box score, (2) triplet preferences around the TSS, (3) hexamer preferences in the regions 11 to 1100, 1101 to 1200, and 1201 to 1300 relative to the TSS, and (4) potential transcription factor binding sites. TSSG is based on the promoter.dat file derived from TFD by Prestridge (1995), whereas TSSW is based on TRANSFAC.

<table>
<thead>
<tr>
<th>Tool</th>
<th>True Positive (out of 24)</th>
<th>False positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audic/Claverie</td>
<td>5 (21%)</td>
<td>33(1/1004bp)</td>
</tr>
<tr>
<td>Autogene</td>
<td>7 (29%)</td>
<td>51 (1/649 bp)</td>
</tr>
<tr>
<td>GeneID/Promoter1.0</td>
<td>10 (42%)</td>
<td>51 (1/649 bp)</td>
</tr>
<tr>
<td>NNPP</td>
<td>13 (54%)</td>
<td>72 (1/460 bp)</td>
</tr>
<tr>
<td>PromFind</td>
<td>7 (29%)</td>
<td>29 (1/1142 bp)</td>
</tr>
<tr>
<td>PromoterScan</td>
<td>3 (13%)</td>
<td>6(1/5520 bp)</td>
</tr>
<tr>
<td>TATA</td>
<td>6 (25%)</td>
<td>47(1/705 bp)</td>
</tr>
<tr>
<td>TSSG and TSSW</td>
<td>10 (42%)</td>
<td>42 (1/789 bp)</td>
</tr>
</tbody>
</table>

**Table 1**: Comparison among existing tools.
3. Present Approach

3.1 Basis of the Work

We have studied the TATA and TATA-less core promoters looking for some discriminating patterns among them. While the tuples of nucleotides of a small length k (we considered k=4) have similar distribution patterns for both TATA and TATA-less promoters (Figure 1) over an extended region around TSS ([-200,+51] segment relative to TSS), we have noticed that some of the tuples have discriminating distributions at the neighborhood of TATA box for TATA and TATA-less promoters (Table 2 and 3).

<table>
<thead>
<tr>
<th>Tuple</th>
<th>Distance from TATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTTC TATA</td>
<td>10 3 6 5 4 2 1</td>
</tr>
<tr>
<td>TATA-less</td>
<td>1 1 0 0 0 2 0</td>
</tr>
<tr>
<td>TCTC TATA</td>
<td>14 6 9 3 4 2 4 2</td>
</tr>
<tr>
<td>TATA-less</td>
<td>1 0 1 0 2 1 2</td>
</tr>
<tr>
<td>AAAAT TATA</td>
<td>0 0 0 4 3 2 3</td>
</tr>
<tr>
<td>TATA-less</td>
<td>6 5 4 3 5 7 9</td>
</tr>
<tr>
<td>GATA TATA</td>
<td>0 0 2 0 2 0 1</td>
</tr>
<tr>
<td>TATA-less</td>
<td>5 0 0 0 2 3 1</td>
</tr>
<tr>
<td>TAAA TATA</td>
<td>0 0 0 4 0 1 4</td>
</tr>
<tr>
<td>TATA-less</td>
<td>3 4 4 3 3 1 3</td>
</tr>
</tbody>
</table>

**Table 2: Tuples vs Frequencies at the Left Neighborhood.**

<table>
<thead>
<tr>
<th>Tuple</th>
<th>Distance from TATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAC TATA</td>
<td>12 18 0 1 2 1 2</td>
</tr>
<tr>
<td>TATA-less</td>
<td>1 0 7 2 2 2 2</td>
</tr>
<tr>
<td>AAAA TATA</td>
<td>23 6 2 2 0 2 4</td>
</tr>
<tr>
<td>TATA-less</td>
<td>2 8 5 4 4 3 2</td>
</tr>
<tr>
<td>AAAG TATA</td>
<td>15 20 2 1 0 1 0</td>
</tr>
<tr>
<td>TATA-less</td>
<td>0 3 0 2 2 3 5</td>
</tr>
<tr>
<td>AATA TATA</td>
<td>60 0 0 0 0 1 2</td>
</tr>
<tr>
<td>TATA-less</td>
<td>10 5 4 3 3 4 3</td>
</tr>
<tr>
<td>TTTT TATA</td>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>TATA-less</td>
<td>6 4 6 7 9 8 4</td>
</tr>
</tbody>
</table>

**Table 3: Tuples vs Frequencies at the Right Neighborhood.**

![Figure 1](image_url)  
**Figure 1.** Distributions of tuples over extended core promoter region covering [-200,+51] relative to TSS

3.2 Promoter Recognition Scheme

Based on the observed discriminating frequency distribution at the neighborhood of TATA we develop two probabilistic models to recognize TATA promoters. The problem of predicting whether a promoter with TATA box is either active or not becomes a binary classification. We can use the following formulation to build a Bayesian classifier for this purpose.

Let C be a random variable for class and it takes values active (A) and inactive (I) indicating whether a TATA box is active and Inactive. Let the string $e_i, e_{i+1}, \ldots, e_{i+k}$ represents k nucleotides upstream and downstream from the TATA box. Then $P(C=A | e_i, e_{i+1}, \ldots, e_{i+k})$ and $P(C=I | e_i, e_{i+1}, \ldots, e_{i+k})$ respectively represent the probability of a TATA being active and inactive in a promoter given their neighboring nucleotides $e_i, e_{i+1}, \ldots, e_{i+k}, \ldots, e_k$. Using Bayes formulas it is rewritten as following.

$$P(C=A | e_i, e_{i+1}, \ldots, e_{i+k}) = \alpha P(C=A) P(e_i, e_{i+1}, \ldots, e_{i+k} | C=A)$$

assuming nucleotides are independent (naive Bayes)

$$P(C=I | e_i, e_{i+1}, \ldots, e_{i+k}) = \alpha P(C=I) P(e_i, e_{i+1}, \ldots, e_{i+k} | C=I)$$

Similarly

$$P(C=A | e_i, e_{i+1}, \ldots, e_{i+k}) = \alpha P(C=A) P(e_i | C=A) P(e_{i+k} | C=I)$$

Where $\alpha = 1/P(C=A) P(C=I)$

From the training set with active and non active TATA box, the conditional probability of $P(e_i | C=A)$ and $P(e_m | C=A)$ for m at positions -k to k BP relative to TATA is computed. When $P(C=A|observation) > P(C=I|observation)$, it is predicted as the TATA box is active, or else (if the condition fails) it is predicted as the TATA box in inactive. This method of prediction using
Bayes method is called maximum a posteriori or a MAP hypothesis.

In the second approach we slid a window of length l at positions–k to k BP relative to TATA and find the probabilities of l-length tuples at the neighborhood of TATA for TATA and TATA-less promoters. For an unknown TATA we examine the neighborhood tuples of the same length and at same positions and using the predetermined probabilities determine whether it’s a promoter element or not.

3.3 Experiments

We apply the naïve Bayes approach for predicting whether a TATA box is active or inactive in a core promoter. For this research work, we obtained data set from PlantProm DB, an annotated non-redundant collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site(s) (TSS) from various plant species. We started the experiment with the offset of 3 nucleotides (-3 to +3 nucleotides ) around a TATA box. From the active and non active TATA box data set, we choose 80% for training and the remaining 20% for testing. We repeated the experiment for 60 times with new set of training and test set maintaining 80, 20 mix.

We use the MAP hypothesis for decision making and computed the mean classification accuracy, true positive and true negative and their standard deviations. Note that false positive is 1-true negative and false negative is 1-true positive.

<table>
<thead>
<tr>
<th>Offset</th>
<th>Overall</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean %</td>
<td>Std %</td>
<td>Accuracy %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean %</td>
</tr>
<tr>
<td>3</td>
<td>53.5</td>
<td>6</td>
<td>75.93</td>
</tr>
<tr>
<td>6</td>
<td>51.34</td>
<td>5.3</td>
<td>69.28</td>
</tr>
<tr>
<td>9</td>
<td>52.40</td>
<td>6.1</td>
<td>64.90</td>
</tr>
<tr>
<td>12</td>
<td>51.48</td>
<td>4.8</td>
<td>64.27</td>
</tr>
<tr>
<td>15</td>
<td>50.94</td>
<td>4.6</td>
<td>63.04</td>
</tr>
</tbody>
</table>

Table 4: % of True and False positives in 1st approach.

The prediction accuracy using naïve Bayes is shown in the graph of Figure 2.

![Figure 2. Prediction accuracy using naïve Bayes.](image)

In the second approach we slid a window of size 4 over immediate left and right of TATA (+/-8 BP relative to TATA) and updated the frequency of the tuple occurring in that window for specific positions. Then we calculated the following probabilities:

i. Probability distribution of tuples of length 4 within a distance of (+/-)8 from TATA boxes.

ii. Probability distribution of tuples of length 4 within a distance of (+/-)8 from all TATA patterns in TATA-less promoters.

Then for unknown TATA we assign scores to the tuples of length 4 located within +/- 8 BP of it using pre-calculated TATA and TATA-less probabilities. We make a predictions based on which score is higher.

Without any bias in the decision making, an implementation of this scheme was giving a very high true positive rate, along with very high the false positive rate showing a bias towards TATA-data. This made the scheme less effective. We noticed that for most of the true positives the difference between TATA and TATA-less scores is much higher than the difference between scores for false positives. In other words, many false positives are barely false positives. We suspect this happened due to insufficient size of training data. We found that an artificially added little bias towards TATA-less scores suppresses those false positives dramatically still recognizing significant number of true positives.

Table 5 shows, the % of true and false positive rate for...
different mix of data. It used 80% of the available data for training and 20% for testing.

<table>
<thead>
<tr>
<th>Offset</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy %</td>
<td>Std %</td>
</tr>
<tr>
<td>5</td>
<td>23.81</td>
<td>7.02</td>
</tr>
<tr>
<td>6</td>
<td>27.85</td>
<td>6.89</td>
</tr>
<tr>
<td>7</td>
<td>31.03</td>
<td>6.17</td>
</tr>
<tr>
<td>8</td>
<td>30.99</td>
<td>6.03</td>
</tr>
</tbody>
</table>

Table 5: % of True and False positives in 2nd approach.

The prediction accuracy of this method is shown in the graph of Figure 3.

![Prediction Accuracy](image)

**Figure 3:** Prediction accuracy using method 2

4. Conclusion

In the present study we investigate how different tuples of small length are distributed over an extended core promoter region covering [-200,+51] segment relative to the transcription start site for both TATA and TATA-less promoters. Experimental results show that frequency distribution of different subsequences over the entire length of the DNA for TATA and TATA-less promoters look almost similar. We suspect that the overall distribution has nothing to do with TATA or TATA-less promoters; rather it represents a general feature of DNA sequences. However, some patterns have relatively high frequency around TATA boxes, but they don’t seem to appear that frequently around TATA patterns in TATA-less promoters and some patterns don’t appear around TATA boxes much, however they appeared more frequently around TATA patterns in TATA-less promoters. This suggests that neighborhood information can be useful in identifying core promoters.

Based on the above observation, we performed statistical learning to predict whether a TATA box is active or inactive in a core promoter sequence. We have applied naïve Bayes with MAP hypothesis for the purpose of prediction. We have divided 80% of the data set (TATA and TATA-less promoters) into training and the rest into test set and perform the prediction and obtained overall classification accuracy, true positive and true negative rate. We repeated the experiment 60 times by generating the training and the test sets maintaining the 80, 20 training test mix. From the 60 runs, we computed the mean and standard deviations. This procedure was carried out to the offset from 3 to 15 in the step of 3. The results are shown in the graph of Figure 2. With the offset 3, the true positive is as high as 76% while the false positive also high 78.3%. When we consider longer neighborhood around TATA box (up to 15 pb in both side) the true positive reduces to 63% along with the false positive, which is 66%. This method is definitely better than any other previous results in predicting true positive. However, it is somewhat disappointing since it has higher false positive rate compared to many other previous works.

As an extension to the naïve Bayes method, we have introduced a method with overlapping tuples of 4 nucleotides that slides in the window -8 to +8 from a TATA box. To improve the results in minimizing false positive, the decision procedure was biased. The best mean value of true positive using the method is 31% while the false positive was reduced to 6.88%. While the true positive rate was comparable with other work, the false positive rate was remarkable compared with other works.

Further study need to be done using other statistical learning techniques so as to improve the overall true positive rate while minimizing the false negative rate.

5. References


Computational Modeling and Simulation of the Immune System

Jugal Kalita, Kaushal Chandrashekar, Ankur Deshmukh, 
Reena Hans AND Priyadarshini Selvam 
Department of Computer Science, University of Colorado, 1420 Austin Bluffs Parkway 
Colorado Springs, Colorado 80917 USA 
kalita@pikespeak.uccs.edu, kchandra@cs.uccs.edu, adeshmu@uccs.edu, 
rhans@cs.uccs.edu, pselvam@cs.uccs.edu

M. Karen Newell 
Department of Biology, University of Colorado, 1420 Austin Bluffs Parkway 
Colorado Springs, Colorado 80917 USA 
mnewell@uccs.edu

Abstract

We have developed a software system called SIMISYS that models and simulates aspects of the human immune system based on the computational framework of cellular automata. We are motivated by the goal of modeling participants in the immune system at the cell level, simulate their interactions and infer overall system behavior. We model tens of thousands of cells as exemplars of the significant players in the functioning of the immune system, and simulate normal and simple disease situations. We present the simulation while in progress with graphical illustration of the participating cells and appropriate graphs. SIMISYS 0.3, the current version of the software, is able to model and simulate the innate and adaptive components of the human immune system. The specific players of the immune system we model are the macrophages, dendritic cells, neutrophils, natural killer cells, B cells, T helper cells, complement proteins and pathogenic bacteria provide communication among the participants in the immune system. In this study we model features of the innate and adaptive immune responses. All immune cells are modeled as classes using object-oriented technology. Cellular interactions are modeled based on the computational paradigm of cellular automata [13]. A graphical user interface is provided so that the user can vary parameters of the simulation. A graphical display, created using the SDL library [10], provides visual images during the simulation.

The organization of this paper is as follows. Section 2 provides an overview of the related research. Section 3 describes the architecture of the software system implemented, and provides details of the object-oriented class structure. In Section 4 we analyze the simulation results. Finally, in Section 5, we discuss future directions for research.

1 Introduction

The immune system is a collection of molecules, cells and organs whose complex interactions form an efficient system that protects the individual from potential harm and outside invaders. Chemical signals provide communication among the participants in the immune system. In this study we model features of the innate and adaptive immune responses. All immune cells are modeled as classes using object-oriented technology. Cellular interactions are modeled based on the computational paradigm of cellular automata [13]. A graphical user interface is provided so that the user can vary parameters of the simulation. A graphical display, created using the SDL library [10], provides visual images during the simulation.

The organization of this paper is as follows. Section 2 provides an overview of the related research. Section 3 describes the architecture of the software system implemented, and provides details of the object-oriented class structure. In Section 4 we analyze the simulation results. Finally, in Section 5, we discuss future directions for research.

2 Related Research

In spite of the enormous complexity of the immune system, three main approaches have been adopted by researchers for computational modeling: Ordinary differential equations (as summarized in [11]), Qualitative, i.e., non-numeric, information for modeling (e.g.,[12]), and Distributed computation using cellu-
lar automata (e.g., [4]). Ordinary differential equations (ODE) have been traditionally used to model complex systems. Perelson and Weisbuch [11] use physical concepts and differential-equations based mathematical methods for modeling immunological problems. They present models for clonal selection and affinity maturation, network models for antibody and B-cell interactions, and autoimmune diseases. However, researchers have also enumerated problems with pure ODE approaches [4]. Some of the problems with ODEs are a) The ODE approach assumes large populations of essentially identical entities, which is not the case with biological cells as each cell has a unique life history that defines its interaction with the environment, b) The ODE approach gives only average behavior of the system, and c) It is difficult to model non-linear behavior.

Cellular automata [13] are discrete dynamical systems whose behavior is completely specified in local terms. They have been widely studied as examples of complex dynamical systems [8], originally as examples of components in a self-reproducing machine [2] and then within the area of artificial life [6]. In a cellular automata model, a uniform grid represents space, with each cell containing a small amount of data. Time advances in discrete steps and simple laws of behavior are used at each step for each cell to compute its new state from that of its neighbors. The behavior of a complex system emerges from simple interactions of simple individuals following simple rules. Cellular automata are sometimes described as counterpart to ordinary or partial differential equations for describing continuous dynamical systems. There have been attempts to simulate aspects of the immune system using cellular automata [1, 5, 7].

3 Implementation Overview

While developing SIMISYS, we have made simplifying assumptions. We follow a systems biology approach [3]. Figure 1 gives the high level view of the immune system according to our model. Details are purposely left out in this figure to keep it simple. Immune cells such as macrophages, dendritic cells, neutrophils, phagocytes, and natural killer (NK) cells are created in the bone marrow. Neutrophils, the most abundant of all white cells, are recruited to the region of a pathogen’s attack in an infected tissue based on the concentration of chemo-attractants. The requirements for activation of antigen-specific lymphocytes, either T or B, are recognition of antigen and other co-stimuli, including cytokines. NK cells need certain cytokines to be activated and kill target cells. They have been implemented as being attracted to the region of infection as a function of the amount of lipopolysaccharide chemical (LPS) produced by the antigen.

The blood vessel is the main port of entry of immune cells into the tissue. Both innate and adaptive players look for a suitable place to exit the blood vessels so that they can enter the lymph node where the lymphocytes become focused to respond to the potential invaders. There is a flow from any location in the tissue to the lymph vessel so that the immune cells responding to the invaders may move to a lymph node. To maintain a continuous movement of immune cells in the tissue, the blood vessels translocate the immune cells from the lymph node back to the tissue. Coordination between the blood vessels and the lymph vessel set up in the simulation has a major role in maintaining flow. For example, antibodies, secreted by the primed B cells, are created in the lymph node and translocated to the tissue. There they opsonize the bacteria. The complement proteins puncture these tagged bacteria in the tissue.

Movement of chemicals such as cytokines in the blood and tissues are modeled using diffusion. A chemical is loaded onto the grid cells at the location of the cell that secretes it. They are diffused through the whole grid depending on their respective breakdown rate and their diffusion constants. This sets up a gradient of chemicals in the tissue allowing for the movement of the immune cells in the tissue based on chemo attractants. This also allows the activation of the appropriate immune cells. In summary, the number of immune cells of various types and their movement are managed by the combined action of the blood and lymph systems and chemical gradients. Figure 4 shows the flow of control between the entities.
and the interactions.

3.1 Software Architecture

The interaction between immune cells and pathogens has been modeled using cellular automata. SIMISYS 0.3 is a complex software system with many interacting components. Figure 2 provides an overview of the software components that constitute SIMISYS.

There are three main components: The Modeler, The Matrix, and The Visualization Engine. The main driver program brings up a graphical user interface (GUI) through which the user inputs parameters to model pathogens and immune cells. The user can vary the parameters describing the cellular players in the system. The user can also provide data to set up the size of the grid and study the impact of changes on the immune response through the GUI. The emphasis of the design has been to create a highly configurable system. Some of the parameters that can be input are given in Table 1. Once the parameters have been specified, control of the system passes to the Modeler. The Modeler reads the values entered and starts the Simulator. The Simulator creates the Matrix, all immune cells, pathogens, blood and lymph vessels, and controls interactions among them. The Matrix models the physical space that the cells occupy. It consists of a 3D grid of cells, where the simulator places all cells and pathogens. Each entity occupies one cell. The matrix also holds chemo-attractants and diffuses them.

The Visualization Module is responsible for display and its simulation. The Data Reader reads the information from the Matrix and provides the information to be displayed to the Display Engine. The Display Engine presents this information to the graphical interface built using SDL which presents a view of the infected tissue or the lymph Matrix depending on the users interest. A separate panel for the display of the statistical results of the system is also provided.

The SIMISYS Immune System simulation is implemented in C and C++. It has a multithreaded architecture based on pthreads[9].

3.2 The Modeler Entities

Some of the C++ objects we use are given in Figure 4. Here are some salient aspects of the object-oriented implementation.

- **Basic Cell:** This is the parent class of all other cell types. Methods common to all cell types are defined in this class. Every cell type is also assigned a corresponding bit signature based on their type. This bit signature is used to differentiate between host cells and invading pathogens and also in antigen presentation.

- **Immune Cell:** This derives from the Basic Cell class and is also the parent class of immune cells. Methods common to immune cell types such as recognizing a bacteria, ingesting the pathogen and antigen presentation are defined in this class.

- **Bacteria:** This derives from Basic Cell type and includes the characteristics of a pathogen such as multiplication, releasing chemicals in their surrounding grid cells and moving to lymph node.
• **Macrophage**: Inherits from Immune Cell class and overloads the methods of immune cell to perform the functions of a macrophage. The initial bit signature is changed after ingesting a bacteria inorder to activate Thelper cells and B cells. The behavior of a macrophage is based on a state variable representing a normal or activated macrophage.

• **Neutrophils**: Inherits from Immune Cell class and contains methods that sense cytokines in blood vessel and move to the site of infection. The life span of objects of this class has a short life span as the real neutrophils.

• **Natural Killer**: Inherits from Immune Cell class. Methods of this class are modified to sense a chemical LPS and produce IFNγ to accelerate other innate players.

• **Thelper cells**: Inherits from Immune Cell. Key functions include recognizing bit signature of an APC, presenting this signature to activate Bcells. Specificity of Tcells has been implemented such that only Tcells a bit signature that complements the bit signature of the antigen are activated.

• **B Cells**: Inherits from Immune Cell and the main function is to secrete antibodies when an antigen with matching bit signature is found. Another important method is to perform clonal selection by which more instances of the specific bit signature are created.

• **Lymph Node, Lymph Vessel and Blood Vessel**: These classes have been created to model the movement of immune cells and pathogens from a site of infection to the lymph node.

• **Antibody and Complement System**: These classes have methods defining the behavior of antibodies tagging the pathogens and the complement proteins destroying them.

3.3 The Matrix

The Matrix represents the physical space we simulate. It is implemented using the following two classes.

• **Grid**: The Grid is composed of an array defined in 3D making up the world of simulation. Each grid cell or box can contain a pointer to an entity and also store information about the current conditions in that cell. Currently we have implemented a 100x100x20 grid. The Grid forms the main section of the tissue where all immune cells move around and interact with each other. Each grid cell maintains a concentration list of all chemicals in it. Examples of chemicals are: LPS (lipopolysaccharide), IFN (Interferon), IL-1 (Interlukin-1), IL-2 (Interlukin-2), TNF (Tumor Necrosis Factor), and IFNγ (Interferon-gamma). Blood and lymph vessels are stationed in the main grid as well. There is a second grid with a size of 30 x 30 x 20 or 18,000 physical cells representing a lymph node.

• **GridWrapper**: This class encapsulates a 3D rectangle of grid cells. It is essential because more than one grid or world is simulated such as a section of infected tissue and a lymph node.

3.3.1 The Visualization Engine

We have developed a tightly integrated visualization engine that is easy to set up and can be adapted to handle introduction of entities in future releases. We also use a graphing package to illustrate the results of the simulation.

3.4 Simulation

The objective of SIMISYS is to simulate various normal and infection scenarios. It is a platform that allows simulation of many different infection and disease scenarios in great detail. The algorithm used for the simulation is given below in a nut shell.
The main driver reads an input file and creates a list of the entities in the simulation. The multithreaded architecture lets one of the linked lists to be selected at random and for each entity of the system, appropriate actions are performed and the entity’s status changed if necessary. The entities which are dead at the end of a simulation cycle are removed from the list and new entities are added if they are created. The dynamics of the system are controlled partly by chemicals and partly by the changes in the life cycle of the entities. One after the other, depending upon which thread of computation gets the control, each linked list of entities goes through the simulation cycle. The display engine also gets the control on a regular basis. Figure 4 gives a flow chart depicting the processes that are modeled and simulated.

4 Results

We discuss four scenarios modeled and simulated using SIMISYS 0.3. They explain the immune response to a simple infection. The numbers displayed in the result graphs of the run of the simulation are the ratios of the number of the entities. On the X-axis, we indicate the number of iterations of the simulation. The change in the number of the associated entities is on the Y axis.

In Scenario 1, the number of Neutrophils increases and subsequently the number of bacteria is decreasing. This is the case of a minor infection in which the neutrophils are effective in fighting against an infection.

In Scenario 2, the neutrophils and NK cells show in the beginning stages of the simulation but are not enough to completely eliminate the bacteria. The chemical released by the bacteria activates the macrophages and the activated macrophages clean up the bacteria.

In Scenario 3, the innate players: neutrophils, macrophages, dendritic cells play their role in the beginning stages and the B cells are also activated to secrete the antibodies. The level of antibodies increases as the simulation progresses which shows an intense cooperation between the innate and adaptive players in fighting against the bacteria.

Situation 4 is a case of bad infection where the neutrophils alone are not able to kill the bacteria. The macrophages enter the battle. The dendrites, which join the innate fighters only when the infection is out of control, also appear. Activated B cells also release antibodies. It takes some time for the antibodies to
tag the bacteria and make their killing by phagocytes easier.

We modified the original implementation to incorporate a simple mode of anthrax infection along with antibiotic treatment. We cannot discuss the anthrax extension due to lack of space. The results of the simulation comply very well with real life.

5 Conclusions

The current model of the innate immune system simulates the self non-self recognition, garbage collection by macrophages, and the role of complement proteins, and the attraction of neutrophils and NK cells to the region of attack. The adaptive part stimulates the activation of T cells, B cells, production of antibodies, and the final action of complement proteins by MAC to kill bacteria. The model also simulates the diffusion of chemicals in the grid and their effect on the functioning of the immune cells. In future versions, we intend to implement further details into all these already implemented classes. We intend to add other immune cells and “organs” such as bone marrow and the liver. We intend to improve the GUI to make it more user friendly. Currently we are working on parallelizing the implementation of SIMISYS on a 32-machine Beowulf cluster so that we can dramatically increase the magnitude of the model.

References


Can new technology impact the Pharmaceutical Industry’s innovation shortfalls?
Layne Los, Virgil Gary Tye

Introduction:
Over the past two decades, the advancement of scientific knowledge and discoveries has expanded at an exponential rate. During the same time frame, pharmaceutical innovation, as measured by new drugs introduced on the market, has slowed. Reasons for this disparity and avenues supporting the translation of scientific breakthroughs to the new technology necessary to speed drug development, are examined herein.

An evolving industry:
Pharmaceutical R&D innovations have declined over the past several years, just as financial pressures and raised expectations (most recently following the sequencing of the human genome) have industry executives looking for ways to improve efficiency. Approaches like Combinatorial Chemistry, High Throughput Screening and the plethora of disease targets gleaned from the human DNA sequence, have not provided the anticipated increase of rapidly developed drugs. In fact, clinical success rates declined from 14% in 1985 to 8% in 2000. (1) This decline is likely due to a number of factors, such as more difficulty measuring clinical endpoints, tighter regulatory requirements, inconsistent Biotech investment, Mega-mergers, and higher economic thresholds (as much as $250M) as Big Pharma companies seek the blockbuster compounds. Corporate culture in large companies reduces efficiency, stifles innovation and creates an environment where risk taking is less likely to be rewarded. While the impact of the mega-mergers on scientific productivity and creativity is hard to measure, merger activity appears to negatively impact productivity while companies reorganize personnel and departments with a new hierarchy in place. Decision making in this environment is often tainted with worry about the impact on job security, and decisions are often reactive, not proactive.

Enter Biotech:
Consolidation in the pharmaceutical industry has changed large corporate structures, thus changing the work environment for the research scientist. Thirty years ago the vast majority of new drugs came from R&D efforts of the big pharmaceutical (Big Pharma) companies, although they weren’t all that big by today’s standards. Biotech’s potential contribution to Big Pharma’s efforts wasn’t evident until critical mass was eventually reached in companies like Amgen, Biogen, Genentech, Chiron and others. Industry investors reaped the benefit of the early success stories and set the stage for business models that became commonplace in the 90’s. Small company cultures developed in the fast paced high science environment dominated by “Gene Jocks” and former Big Pharma executives looking for a more meaningful role in decision-making. Biotech companies became Big Pharma’s source of developable drugs that could augment dwindling drug development pipelines. The smaller company’s survival often depends on how quickly new technology could advance programs. The innovation and rapid pace had a natural regulator built into the system. An often resource-constrained management team needed to develop the necessary symbiosis with Big Pharma. Even well-funded Biotech companies lack the financial and logistical resources to develop drugs beyond Phase I. Finances and the need for innovation drove collaborations between the progressive Biotech companies and the more conservative Big Pharma companies.
The Contract Research Organization (CRO) concept:
The level of expertise and specialization of CROs has significantly increased over the past decade. Industry scientists understand the changing role of the CRO and value their contribution to the Drug Development process. Many large, well-funded R&D operations send their most challenging projects to CROs because of very specialized skill sets developed from the vast range of technical programs CROs experience. Generally there are 3 types (or levels) of CRO’s that fill various needs in the Pharmaceutical R&D. Large, multinational organizations such as Quintiles, Covance, MDS, and Huntingdon, are able to do “most everything”. Middle tier labs, like Wil, MPI and Purdue, have grown (diversified) from their original specializations such as Toxicology, or Bioanalytical. These companies are likely to add other services, as they become financially attractive additions. Thirdly, there are specialty labs with unique skill sets that have roots in the founder’s scientific expertise, for example; ion channels (ChanTest); hemotoxicity (HemoGenix); or cell-based toxicity (CeeTox) and predictive ADME (ADMETRx). It is the last group that NextGen Pharma Technologies believes will continue to be a primary source of drug discovery tools utilizing new technology methods.

Challenges Faced By Small Innovative Companies
Challenges for young innovative companies are considerable. Resources issues are the primary reason these companies have difficulty translating new technology into a marketable service.
- Bootstrap; or self-funded companies retain control of their companies but are often undercapitalized which limits growth.
- Few scientists have business training. Depending on the entrepreneurs’ inclination, this may slow growth or make the organization prone to mistakes that can cripple a company especially in the early stages.
- Venture capitalists money is available to some - but the VC’s business-focused objectives are sometimes incompatible with long-term success of the company. VCs commonly require the entrepreneur to relinquish control of the company as a condition of funding, which may be unacceptable.

SBIR Grants are an excellent funding source for new product development young companies especially in light of recent improvements in funding options. However, reliance on SBIR grants has significant drawbacks primarily. Long cycle time - submissions to funding, and the submission process is daunting. There are ways to improve SBIR success rates, such as delegating the application process to consultants specializing in guiding entrepreneurs thru the process.

FDA Perspective
The recent position paper published by FDA in March 2004, “Innovation/Stagnation - Challenge and Opportunity on the Critical Path to New Medical Products”, gives the FDA’s perspective on reasons new technology has been slow to improve the drug development process. The paper makes recommendations and outlines initiatives to address the problem. This author believes the FDA’s new initiatives may facilitate the implementation of scientific breakthroughs. Most importantly, the agency has made new technology a focal point, opening the door for an expanded dialogue addressing the agency’s role in slowing the implementation of new technology (which was not addressed in the paper). For example, companies that utilize new technologies may fear they will be burdened with performing complex and costly validation studies. Over the past decade this author has witnessed a reluctance of some pharmaceutical executives to use new technology or animal models for fear it will open Pandora’s box, requiring the company to allocate significant resources to answer endless questions. An FDA reviewer’s scientific curiosity and/or, perhaps, inexperience have
contributed to an impression that the application of new technology becomes a research project funded by the drug company. Perhaps the best example involves FDA’s recommendation in the late 90’s that companies use alternative animal models, such as transgenic mice, for Carcinogenicity assessments, instead of the tried and true (yet seriously flawed) 2-year rat bioassay. The enormous amount of uncertainty associated with interpretation of these results did not encourage use of the transgenic model. The potential for a positive result that could kill a compound was especially worrisome. If you wanted to dispute the result, a 2-year Carcinogenicity study was needed anyway. Even if the 2-year study result contradicted the transgenic mice result, defending your compound remained a arduous task. Although the frequency of these hurdles is not well documented, financial obstacles such as this are formidable for any drug company. There is a reluctance to generate results that lack precedent and may not be easily explained, fearing that the FDA will require further costly and time-consuming studies to validate or disprove the initial results.

**A new FDA initiative:**
The FDA has begun an initiative to identify academic researchers and new innovative technologies that improve/facilitate drug discovery and development. Focus on collaborative efforts to development “Critical Path Opportunities List” for emerging technologies that advance and modernizes the process. Areas identified include:

- Predictive Toxicology
- Proteomics and toxicogenomics
- Cardiovascular assessment for heart rhythm abnormalities
- Imaging technologies for Neuropsychiatric diseases that assess drug Pharmacodynamics
- Clinical study design and patient-driven outcomes research
- Knowledge-based drug development
- Model-based clinical trial simulation software
- Pharmacogenomic-based Biomarkers for drug effectiveness
- Stem cell characterization

**What works:**
New technologies have the potential to significantly improve the drug development decision-making process. One recent example is Predictive ADME, now widely implemented in early drug discovery. A project team armed with *In Vitro* solubility, permeability, metabolic stability and drug transporter data will have a higher probability of choosing a successful drug candidate. The availability of these data is likely responsible for the reduction in clinical failures due to ADME from 40% to less than 15% over the past 5 years. Similar gains are anticipated in Predictive Toxicology, Toxicogenomic, and Biomarkers as the technology and databases are further developed and validated.

NextGen Pharma Technologies has created business models and relationships that support and nurture entrepreneurs by providing contract sales efforts and business advice. The primary focus is to provide a revenue stream through sales and marketing efforts early in the company’s life cycle. If a company can rapidly become self-sustaining, it allows them more freedom and better options to grow their business. Because cash flow is key to the success of any company, sales and marketing is the top priority.

Strength in numbers is an old adage that rings true. NextGen can provide Pharma executives interested in new technologies, access to a dozen or more young companies with proven technology. NextGen’s third party endorsement and independent assessment are more credible. Third party feedback helps steer a company’s early R&D emphasis for new products. Quality and turnaround
issues can be addressed independently, which is an optimal situation for quality evaluation. Collaborations between NextGen companies promote the collective stability of the group. While NextGen is not a source of investment dollars, it provides customer feedback and contract revenue much earlier and more cost effectively than internal efforts. NextGen’s clients share travel and promotional expenses, which also conserves cash.

**Examples:**

**CeeTox and ADMETRx** are unique examples of technology companies developed from expertise gained while employed in Big Pharma. Pfizer’s acquisition of Pharmacia allowed key scientists to move fully functional departments in Pharmacia’s discovery organization. In this example, state and local Kalamazoo officials had the insight too provide incubators lab space that would allow these scientists to remain in the area. Seed funding was also provided on a limited basis.

**CeeTox** – developed a Predictive *In Vitro* Toxicity screen using known biochemical end points that assess liver cell cultures health while exposed to log scale drug concentrations. The technology was validated with *In Vivo* rat toxicity studies (3). CeeTox rank orders compounds while predicting with 85% accuracy the blood concentrations that will cause toxicity.

**ADMETRx** – the mission at Pharmacia was to support medicinal Chemists’ efforts to design in “drugable” properties into lead optimization candidates. The team at ADMETRx was able to take what they learned over 20 years (4,5,6) and apply a refined decision tree, experimental design, and local vs global modeling approach to provide precise predictive ADME models.

**Source Precision Medicine** – was started by former Amgen executives at the company’s Boulder, Colorado site. Source focused on further refinement of QPCR technology to the point where it’s reliability and precision could discern changes in inflammation gene expression that correlated with the disease severity and drug response (7,8). Considered rheumatoid arthritis biomarkers they reduce the cost and time to conduct clinical RA trials.

**ChanTest** – was at the right place at the right time. Academic researchers who discovered and published on potentially fatal drug interactions provided an answer to patient deaths caused by Seldane induced cardiac electrophysiology abnormalities (9). ChanTest grew a company with revenue generated from laboratory screening studies that became require by FDA. Drugs that are about to be administered to humans for the first time are tested for their cardiotoxicity liability. ChanTest continues to expand on its definitive regulatory study capabilities by adding new higher through put screening technology based on market demands.

**HemoGenix** – started when an academic researcher supporting bone marrow transplantation unit in a South Carolina VA hospital, began working on a better way to automate the labor-intensive colony-forming assay. Several NIH and SBIR grants earmarked for validating the new technology to decades of historical data from colony forming assays developed in the 1960’s bolstered HemoGenix initial modest funding. HALO, the new assay based on cell proliferation, uses 96 well format with a luminescence output that dramatically expands the ability to measure drug effects on bone marrow and will be approved as a quality control measure of bone marrow transplantation (10,11).
**InVivoMetrics** – is currently a University based, virtual company using MRI as a non-invasive imaging tool in drug development. The company develops imaging coils for specific applications yielding high-resolution images in laboratory animals. The services provided to researchers help measuring drug effects on disease states and prevention.

**NextGen Pharma Technologies** has coordinated a collaboration that yields a unique telemetry data collection capability coupled with unique EEG data evaluation software, in dogs and rats. The software generates a unique drug signature map, which point toward the drugs main pharmacological effects. The technology has been used to track a drug’s effects on brain wave activity.

**Conclusions:**
Translations of scientific discoveries into technologies that improve the drug development process are fraught with difficulties and barriers. This is a reflection of the aversion toward risk due to the nature of the pharmaceutical industry. Traditional safety assessment tools have not changed in decades. Efforts to supplant the decades of industry experience and historical control data with new technology seem insurmountable. Business conditions such as mega-mergers and unreliable cycles of Biotech funding have created periods of instability.
Opportunities have emerged from technology developed in Big Pharma and Biotech companies. Big Pharma mergers and small technology-based company bankruptcies have provided the core of Contract Research Organizations. In addition, industry and academic scientists with the fortitude to obtain NIH and SBIR grant money, have also developed new technologies that aid pharmaceutical research. These scenarios have had a significant impact on the availability of new technology on a fee for service basis.
The future of the pharmaceutical industry and its ability to make use of scientific breakthroughs to aid research looks brighter than it did several years ago. The recent FDA position paper challenges the industry to look ahead and take steps that will reverse the trend of reduced efficiency. This author believes the paper falls short of an honest and critical assessment of the agency’s own role in stifling the drug development process. However, this may mark the beginning of a new era of FDA sponsored support of new technology.
References:

1) “Innovation/Stagnation - Challenge and Opportunity on the Critical Path to New Medical Products”. US Department of Health and Human services - Food and Drug Administration, March, 2004


3) “Strategic Importance of Research Support through Pathology” Gary L Cockerell, James M. McKim, Steven L. Vonderfecht. Toxicologic Pathology, Vol 30 No. 1, PP 4-7, 2002

4) “Advancing In Vitro ADME/Tox”. Drug discovery & Development Vol 5 No 4 April 2004


7) High Precision Gene Expression Analysis Distinguishes Therapeutic Response in Rheumatoid Arthritis (RA), John Cheronis, Victor Tryon, David Trolinger, Vicki Abbas, Richard Gerren, Nancy Nicholls, Brian Kotzie, Michael Schiff, Gerd Rudiger Burmester, Hiroaki Mitsuya, Tatsuru Kawaguchi, Ken-mei Sakata, Micheal Bevilacqua. ACR annual meeting, New Orleans Oct 2002


Growing the Family Tree:  
The Power of DNA in Reconstructing Family Relationships

Luke A. D. Hutchison*  Natalie M. Myres†  Scott R. Woodward‡

Sorenson Molecular Genealogy Foundation (www.smgf.org)  
2511 South West Temple, Salt Lake City, Utah 84115, USA

Abstract

The Sorenson Molecular Genealogy Foundation is building the world’s largest database of correlated genetic and genealogical information to enable genealogical research to be performed using DNA analysis techniques. DNA samples with associated 4-generation pedigree charts have so far been collected from approximately 40,000 volunteers. Up to 170 regions of DNA are currently analyzed for each individual, and the corresponding pedigree chart is extended as far as genealogical databases allow, to currently include over 700,000 ancestral records. By combining these two sets of correlated data on an unprecedented scale, we are enabling progress for the first time into the new field of “molecular genealogy.”

Molecular genealogy is the application of DNA analysis techniques and statistical population genetics to the task of reconstructing unknown genealogies from the genetic and genealogical information of living individuals. We address aspects of using DNA for genealogical research, including those of identification and differentiation of populations (with population boundaries defined not just by factors of demographic separation, but also by time periods), differences in inheritance models of the various types of genetic data, clustering, statistical reconstruction of ancestral trees, inference of ancestral genetic signatures, and inference of surname based on paternal-line DNA.

1 Introduction

Every living individual carries within themselves a combination of the genetic signatures of their ancestors. This unique combination of signatures forms the individual’s unique genetic identity, which is subsequently passed on to become a constituent part of succeeding generations. We are thus intrinsically linked to, and part of, our forebears and our descendants. Truly, in the words of Donne, “No man is an island, entire of itself; every man is a piece of the continent, a part of the main . . . any man’s death diminishes me, because I am involved in mankind.” [1].

The vast majority of our DNA is identical to that of all others in the human race. It is this pattern, common to all human life, that identifies us as human. And yet, almost paradoxically, the small differences between genetic signatures give us identity as individuals. The number of genetic markers that differ between humans is disproportionately small compared to the total size of the human genome (the genetic “blueprint” of each human being). However, the total number of differences between any two humans is numerically large enough that each individual is unique among all other individuals who have ever lived. The regions of DNA that differ between individuals, known as polymorphic sites, give us a unique identity and a place in the human family tree.

Molecular (or genetic) genealogy is the application of DNA analysis techniques and statistical population genetics to the task of reconstructing unknown genealogies from the genetic and genealogical information of living individuals. The purpose of molecular genealogy is to supplement, not supplant, traditional techniques for genealogical research. The types of answers that may be provided by molecular genealogy include the derivation of populations of origin of unknown ancestors at genealogical “walls;” the reconstruction of ancestral genotypes (genetic signatures) from the genotypes of living descendants; the quantification of relatedness and possible kinship of two individuals; the inference of surnames on paternal lines in patronymic lineages; the investigation of possible non-paternities or adoptions; and, ultimately, the reconstruction of unknown pedigrees, or the tying of living individuals to specific previously-unknown ancestors.

The Sorenson Molecular Genealogy Foundation (“SMGF,” www.smgf.org) is building the world’s largest database of correlated genetic and genealogical information, to enable genealogical research to be performed using DNA analysis techniques. Currently, DNA samples with corresponding 4-generation pedigree charts have been

*Director of Bioinformatics, SMGF – luke@smgf.org  †Co-Principal Investigator, SMGF – natalie@smgf.org  ‡Principal Investigator, SMGF – scott@smgf.org
collected from approximately 40,000 volunteers. The DNA for each sample is analyzed at up to 170 locations across the genome, and the corresponding pedigree chart is extended as far as genealogical databases allow, to include over 700,000 ancestral records. The combination of these two types of correlated data on an unprecedented scale presents rich opportunities for analysis, and uncovers new, challenging problems by enabling the rst real large-scale exploratory research into the eld of molecular genealogy.

2 Types of Genetic Data

2.1 Sequence Data

DNA sequences are the most fundamental form of genetic information. The four nucleotides, abbreviated A, G, C and T, are the atomic units of a DNA sequence. Cells in the body contain four billion pairs of nucleotides (referred to as bases) that uniquely identify the individual, and that completely specify the structure and function of the entire organism. DNA sequences differ between individuals, predominantly because of the genetic processes of mutation and recombination. Algorithms exist for nding the "genetic distance" between the sequences of two individuals, or the number of edit operations (insertions, deletions and substitutions) needed to convert one sequence into the other [11]. While complete DNA sequence data can be used to derive all other genetic data, currently it is prohibitively expensive and time-consuming to obtain substantial sequence data for large numbers of individuals.

2.2 SNPs

Single Nucleotide Polymorphisms (SNPs, pronounced "snips") are single-base mutations in a DNA sequence where one base changes to another (Figure 1). These tend to be rare events (in some cases, unique events in the history of the human race), with mutation rates estimated at around 175 total SNP mutations per individual per generation, or .000002% per base per generation [8]. SNPs thus allow for the tracing of extremely deep-rooted pedigrees. SNPs are more useful for anthropological studies than genealogical studies because of their typically low mutation rate. Considering multiple SNPs together provides the ability to more accurately pinpoint the actual time of divergence of two ancient lineages, and allows for non-unique-event SNPs to be identified.

The Y Chromosome Consortium [13] has identified a set of SNPs useful in classifying males into populations of origin. They present a decision tree for hierarchically classifying individuals into major clades or lineage forks, then into specific haplogroups, or subgroups of more closely-related individuals within each clade. The hierarchical designation appears to map reasonably closely to the demographics of known ancestral populations of tested individuals. It is worth noting though that these are paternal-lineage populations, because the SNPs used are all on the Y chromosome. Paternal-lineage populations have different properties than do traditional populations, as will be explained in Section 4.1.

2.3 STRs / Microsatellite Loci

A short tandem repeat (STR) or microsatellite locus is a region (or locus) of DNA in which a repeat unit, in the form of a specific sequence of bases, is repeated a number of times (Figure 2). The repeat region is amplified (copied millions of times) using PCR (Polymerase Chain Reaction), and is then genotyped to determine the number of repeat units at each locus for each DNA sample. The number of repeats, or allele value, at a particular marker or locus on a chromosome is passed down from parent(s) to child unchanged, unless there is a mutation, which will usually make the region longer or shorter by one complete repeat unit. STRs tend to have much higher mutation rates than SNPs (estimated at around 0.3% per locus per generation [7]), meaning they are much more useful on a genealogical timescale.

Of the different types of genetic data, the most cost-effective to obtain in large quantities is currently STR data. (Techniques for detection of SNPs and sequencing of DNA on a large scale are rapidly improving however.) Consequently, most current research in molecular genealogy primarily employs STR data.

3 Genetic Mechanisms Affecting Molecular Genealogy

3.1 Mutation Models and Mutation Rates

In general, genetic variation between generations results from the genetic processes of recombination and mutation, and may happen at the individual-base level, or may affect multiple bases (for example, in the case of entire STR repeat units being inserted or deleted).

While the occurrence of mutation is taken for granted, exact models that describe the mutation process are not known, particularly in the case of STRs. Some of this difficulty arises from the low probability of actually observing a mutation at a specific locus in any given generation, and from the size of the average generation gap in humans. Models have thus been proposed to approximate the actual behavior of a locus over many subsequent generations, including the in nite alleles and stepwise mutation models [6]. It is generally agreed that these models over-
simplify the actual process of mutation, although they do provide useful tools for analysis of patterns of mutation under certain limited conditions.

Mutation rates have only been estimated approximately, and for small numbers of loci, due to these difficulties in actually observing mutations [5, 7], and because of the time and cost currently involved in determining the genotypes of large numbers of individuals. The rates which have been determined by observational studies appear to vary significantly between loci, meaning that a single average mutation rate cannot meaningfully be applied to all STR loci for most purposes.

3.2 IBD vs. IBS

If an allele (genetic marker value) at a specific locus is passed down from an ancestor to two descendants unchanged, it is said that the two descendants are identical by descent (IBD). If the two descendants mismatch due to one or more mutations, then the descendants are said to be different by state (DBS). However the two lineages may separately mutate away from the original allele, and then eventually randomly mutate again to a matching configuration. This is known as identical by state (IBS) (see Figure 3). IBD matches occur over relatively short timescales (as no mutation has been observed on either lineage); DBS mismatches typically occur over longer timescales; and IBS matches typically occur over much longer timescales (because multiple mutations are observed). IBS matches can be problematic, because if treated as IBD matches, they would imply a much shorter time to most recent common ancestor (TM RCA) than a true IBD match. Analyzing several loci together can help discern IBS matches, because if a large proportion of loci match between two individuals, it is much more likely that the matches are IBD than IBS.

The infinite alleles model, mentioned above, assumes that every mutation produces a new, globally-unique allele, disallowing IBS matches. This serves to simplify many mathematical analyses, but does not capture the reality that IBS matches occur a great deal in nature. IBS matches are particularly a problem when the mutation rates of loci under consideration are very different.

4 Genetic Inheritance Models

4.1 Y Chromosome (Ycs) DNA

The Y chromosome, possessed only by males, is passed down from father to son mostly unchanged. The majority of the Y chromosome is formed of non-recombining, haploid (non-paired) DNA, meaning the changes that arise in the Y chromosome are primarily due to mutation. Typically, the Ycs markers that are used for molecular genealogy are STR loci in the non-recombining (NRY) region, with an average
mutation rate of approximately 0.3% per locus per generation [7].

The inheritance model of the Y chromosome is immediately useful to genealogists, because it follows the same inheritance pattern as that of surnames in many western (and even non-western) societies. Thus, there is a correlation between observed Y chromosome genotypes and surnames. This is not a 1-to-1 correspondence, because of adoption, non-paternity, multiple origins for the same name, mutation, etc., but a fuzzy search against a database of surname-labeled Y chromosome genotypes nevertheless provides a useful way of finding possible family names beyond these events on paternal lineages. It also provides a means to identify others who share common biological ancestors on the paternal line where there was an unknown biological relationship, helping genealogical researchers who were unaware that they were biologically related to find each other. On a coarser scale, there is a correspondence between DNA patterns found in the Ycs and various world populations, which can allow researchers to trace the population of origin of a paternal-line ancestor.

It should be noted that the definition of population or cluster is somewhat unusual when dealing with the Y chromosome, because we are considering non-recombining paternal-line DNA. The characteristics of paternally-related populations are different from those of populations defined by recombining DNA (which produce the “traditional” definition of a population). For example, multiple unrelated lineages (paternal populations) can coexist in a common geographical location for an indefinite period of time without direct genetic interaction. Populations defined by non-recombining DNA are immune to traditional population-genetic forces that are caused by recombination, such as inbreeding. Paternal populations are also not affected by population growth effects in the same way as traditionally-defined populations, such as in the case of a historic contraction in population size. Population contraction affects a haploid population in the same way as a slow population expansion (possibly with genetic drift over time), followed by a rapid expansion. From the point of view of present-day genetic analysis, it is as if branches of the Ycs tree that did not make it through the population contraction never existed, because there is no further trace of the Ycs of specific paternal lines that ended at some point in history.

In this light, it makes sense to define a paternally-related population or cluster as a group of individuals who share a common paternal ancestor recently enough to match IBD at a significant number of loci, or as a group of lineages that descended from a common ancestor and whose living descendants are still genetically similar to their common ancestor. This definition of paternally-related populations necessarily impacts any inferences about paternal population structure that are made from the Ycs of living descendants. Paternally-related populations are specifically defined by common ancestry, and thus are only indirectly correlated with geographical origins (due to the physical location of the common ancestor).

Several methods for visualization of relatedness of entire populations have been developed. Many of these methods take a matrix of pairwise relatedness measures as input. Cladograms (branching tree-like diagrams) were extensively used in the past to visualize relatedness among individuals; they have been superceded by median networks (diagrams that may possess loops) and other visualization methods, due to the fact that cladograms yield results that are in general not theoretically sound (they do not capture the true nature of relatedness between individuals in different branches). Attempts are also often made to reconstruct the actual Ycs inheritance tree (or set of paternal lines descending from a common ancestor), using phylogeny (tree-building) algorithms such as those provided by PHYLIP [4]. These algorithms employ a form of heuristic random search of all possible lineage trees under specific phylogeny criteria, and yield an approximate solution. The best phylogeny for a dataset usually cannot be determined, because the total number of trees that may be reconstructed for a dataset of a given size scales exponentially with the number of individuals, quickly rendering the problem intractable (uncomputable) for moderately-sized datasets. Unfortunately, while the output of a phylogeny algorithm is generally accepted as authoritative, the search space is so large, and the pairwise-distance data often so internally inconsistent (due to IBS matches and limited numbers of actual loci in the genotypes) that different phylogeny algorithms and different runs of the same algorithm almost always give different results. For these reasons, phylogenies generated by current algorithms should be treated as informative but not authoritative.

4.2 Mitochondrial DNA (mtDNA)

Like the Y chromosome, mitochondrial DNA (mtDNA) is haploid (non-paired). It is present in the mitochondria, or energy-producing units of the cell, rather than in the nucleus. There are typically hundreds of mitochondria per cell, and multiple mtDNA molecules per mitochondria. The mother’s mitochondria are those present in the zygote, or first cell of a new human being, and thus a mother passes her mtDNA to her sons and daughters. Her sons, however, will not pass their mtDNA to the next generation. Thus the mtDNA may be thought of as almost exclusively maternal-line DNA. Most of the observations made above for Ycs DNA and paternally-related populations are also true for mtDNA and maternally-related populations, because mtDNA is essentially inherited along the maternal line.
4.3 Autosomal DNA

Autosomal DNA is the diploid (paired-chromosome) DNA that forms the vast majority of the DNA in most human cells. Each of the two alleles at a specific locus on an autosomal chromosome of one parent has a 50% chance of being passed on to each child, meaning that on average, a specific allele is passed on to half of the children. Additionally, autosomal DNA recombines, meaning that at each generation, sections of DNA are exchanged between pairs of corresponding (homologous) chromosomes received from the two parents.

If there is a low probability of a recombination between two or more loci, then they have a high probability of being inherited together by successive generations, and the loci are said to be in linkage disequilibrium (LD), or simply linked. Loci may be linked if they are located physically close together on the chromosome, or because there are few potential recombination sites between the loci. It is even possible to observe statistical correlation or linkage by association between distant sites, meaning that specific combinations of alleles at the loci occur together much more frequently than can be accounted for by chance. Groups of alleles on a single chromosome at loci that are in disequilibrium are called haplotypes. Haplotypes take a much greater range of possible combined forms than the individual loci they are comprised of, meaning they are more specific than individual loci, and are therefore more useful for genealogical purposes. A single STR locus may have ten possible allele values, meaning everybody in the population falls into one of ten categories, resulting in a moderately high chance of IBS match between two random individuals. A 3-locus haplotype, however, may have over a thousand possible configurations of alleles, resulting in an increase in specificity and a decrease in likelihood of IBS match.

When analyzing multi-locus genotypes, it is impossible to determine which chromosome of a pair a specific allele came from – the data is said to be unphased. The problem of determining which alleles at each locus of a set of linked diploid loci are physically located on the same chromosome is known as haplotyping or determining phase (Figure 4(a)). For example, for a set of three linked autosomal loci, we have $3 \times 2 = 6$ alleles in the unphased genotype, yielding a maximum of $2^4 = 8$ possible assignments of alleles to specific chromosomes, or $2^{3-1} = 4$ possible phases when not distinguishing between the chromosomes. Depending on the allele values, some of these alignments or phases may be identical to each other, due to homozygosity (where the two alleles at a locus are identical).

If the genotypes of the parents are unknown, it is not possible to determine which allele in the child came from the father and which allele came from the mother. Additionally, only one of the two alleles at each locus is passed down from each parent to each child. When three or more siblings’ genotypes are known, it may be possible to reconstruct the two parent genotypes unambiguously, but it is not possible to determine which genotype corresponds to which parent (the mother or father) without genotypes of extended families. Once phase is set in a child, the assignment of the two resulting haplotypes to the correct parent of origin is important, in order to be able to propagate haplotypes back through the genealogy, to infer ancestral types (Figure 4(b)).

Autosomal loci that are unphased and unlinked (and therefore not able to be haplotyped) are very difficult to trace genealogically without knowing the genotypes of large numbers of individuals in an extended family group, because a specific allele could have come from either parent at each generation, and only one of each parent’s two alleles was passed on to each child. However, unlinked autosomal loci can still be used for genealogical purposes, by clustering together similar individuals, and then looking for patterns in the geographic origin of the known ancestors of the individuals that fell into the same cluster. Populations in general have a distribution of alleles that is distinct from that of other populations.

One algorithm that does a reasonable job of clustering individuals based on unlinked autosomal loci, known as STRUCTURE, uses a Markov Chain Monte Carlo (MCMC) algorithm to iteratively improve cluster membership probabilities until a reasonable solution is found [10]. STRUCTURE has been used to cluster many autosomal datasets.
4.4 X Chromosome (Xcs) DNA

The X chromosome has a very interesting inheritance model: because each male has an X-Y combination of sex chromosomes and each female has an X-X, males of necessity received their Y chromosome from their father, and received one of their mother’s two X chromosomes. Females received one X from their mother and one from their father. This is useful in haplotyping the X chromosome in females – as the X chromosome is haploid in males and diploid in females, it is possible to always unambiguously set phase in the genotypes of any mother-son or father-daughter pair. By creating a dataset of phase-known females mixed with phase-unknown females, we can estimate how well any given haplotyping algorithm performs in haplotyping the Xcs in a large female population: the accuracy with which phase was determined for phase-known females gives an estimate of performance on phase-unknown data.

This is a good model for estimating performance of haplotyping algorithms on autosomal data, since Xcs STRs are believed to have genetic properties in females that are similar to those of autosomal STR loci.

Haplotyping has so far proven to be a difficult problem, although several researchers have created tools that can successfully reconstruct a large proportion of haplotypes from a set of random simulated genotypes [2, 12]. Determining phase for autosomal loci is difficult when the relationship of individuals is not known, because analysis can only be performed on a population level. It is hard to check the validity of haplotyping results, because the haplotype phase was unknown to start with. In order to test the validity of phase reconstruction, haplotyping algorithms are typically tested with data that is simulated and therefore of known phase. In our experience, these algorithms do not work nearly as well as claimed when they are applied to real, phase-known data, such as the Xcs data we have obtained from known father/son and mother/daughter pairs in our database. Haplotype algorithms can also be very slow to run. At SMGF, we have created a new haplotyping algorithm that sets phase in a population of haplotypes with an accuracy that is close to that of the current best algorithm, PHASE v2, yet runs several orders of magnitude faster. This algorithm will be described in a future publication. Our dataset of 220 phase-known individuals (combined with several thousand phase-unknown individuals), derived from real data, will also be of interest to those working on the haplotyping problem.

4.5 Comparison of Inheritance Patterns

It is interesting to compare the modes of inheritance of the various human chromosomes in the context of genealogical reconstruction. The chromosomal inheritance patterns have different characteristics depending on whether the inheritance is considered forwards or backwards through time.

The Y chromosome, for example, may be inherited by any number of sons at each generation, yielding a paternal tree relationship when viewed forwards through time. However, each son received his Y chromosome from exactly one father, yielding a single paternal lineage when viewed backwards through time. Mitochondrial DNA has very similar inheritance patterns on the maternal line, producing a maternal tree and maternal lineage if viewed forwards and backwards respectively, except that the “maternal tree” also has male leaf nodes (sons) connected to many of the female nodes in the tree.

Autosomal alleles follow a zig-zag pattern (single-path random walk) back through time, since they could have come from either parent at each generation. The number of possible ancestors that any given autosomal allele could have come from at the nth generation is 2^n. Interestingly, if one traces a pedigree chart far enough back, the same ancestor begins to appear on multiple branches of the pedigree: the pedigree actually coalesces. Even further back, the founding ancestors of the human race would appear on every branch of the pedigree – or, if the pedigree chart were drawn such that coalescing ancestors were drawn once, the chart would diverge and then converge again (this is effectively what is known in discrete mathematics as a lattice, a specific form of partial ordering). Also, at some recent point in human anthropological history, large proportions of those living today shared almost all of their ancestors [9] due to extreme coalescence of ancestral lines (Figure 5).

Looking forward through time, autosomal alleles are potentially inherited by multiple children at any generation, so a single allele follows a path that resembles a lightning bolt (i.e. the forward-inheritance mechanism is a branching random walk).

The X chromosome, however, has the most intriguing mode of inheritance. When viewed forward through time, each male may pass their X chromosome to zero or more females (and exactly zero males), and each female may pass their X chromosome to zero or more children (male
or female). Looking backward through time, the number of potential ancestors that could have been the source of any given allele on the X chromosome at generation $n$ back grows as the sequence $F_n = 1, 1, 2, 3, 5, 8, 13, \ldots$. This will be familiar to many as the Fibonacci Sequence, whose ratio of successive terms converges upon the Golden Ratio $\phi = 1.618$ (Figure 6).

![Diagram](image)

**Figure 6:** The number of ancestors at generation $n$ from whom a living individual may have received an X chromosome allele is $F_n$, the $n$th term of the Fibonacci Sequence. The ratio of successive terms in the Fibonacci sequence converges on the Golden Ratio $\phi = 1.618$.

5 Importance of Genealogical Data

The true importance of the SMGF database for molecular genealogy lies in the genealogical data that accompanies each of the 40,000 genotypes, which currently totals over 700,000 ancestral records. The presence of comprehensive genealogical data, polished by qualified genealogists, for every DNA sample in the database, allows for an entire dimension of analysis that is not possible using the genetic data alone. The combination of genetic and genealogical data present in the SMGF database is unprecedented on this scale.

In particular, it is important that identity linking is performed as accurately and thoroughly as possible. To statistically reconstruct genotypes of ancestors, we need to know the DNA of as many living descendants of that ancestor as possible. If an ancestor is present in the unlinked pedigrees of several different individuals, then the ancestor has several different identities in the database, and there is significantly less information available to infer the ancestral genotype. Conversely, the more correct identity links that are made for a common ancestor of living individuals, the stronger the inferences that can be made as to the ancestral genotype, since DNA from the ancestor is likely to have made its way to multiple living descendants at a higher relative frequency than that found at random in the population. Without these links, at least for autosomal DNA, each allele is equally likely to have come from any one of the ancestors at a specific generation. Thus the key to verifiable molecular genealogy, particularly for recombining DNA, lies in accurate identity linking. It is very likely that better linking technology would result in the identification of further identity links between many of the 700,000 ancestral records in the SMGF database.

This raises issues of data accuracy and datafield normalization – much genealogical data is incomplete, incorrect, or inconsistent between different sources. The error rate and incompleteness rate increases the further back the genealogy is traced. However, there is a significant percentage of available genealogy that is certainly correct; we minimize initial error as much as possible by employing proficient genealogists and by drawing from the best data sources, and then it is our goal to identify remaining inconsistencies in the genealogical data by consulting the DNA.

It is interesting that genetics can serve as a verification of genealogy, and vice versa. Genealogy can also serve as a prior for genetics-based pedigree reconstruction, with the effect of reducing the total problem space, and of detecting and correcting errors by providing informational redundancy (as with error correcting codes in data transmission).

6 Population Genetics

Statistical population genetics provides many important clues and analysis techniques to achieve the goals of molecular genealogy [3]. In particular, quantification of various genetic and population parameters (such as gene diversity, locus homogeneity, kinship coefficients, linkage disequilibrium measures and average time to most recent common ancestor) can aid in understanding a population’s genetic history.

Much of statistical population genetics relies upon a simplification of population dynamics, known as Hardy-Weinberg Equilibrium (HWE). A population is in HWE if the population is infinitely large, there is no migration to or from the population, all members of the population reproduce, all mating is random, everyone produces the same number of offspring, and neither mutation nor natural selection occur. There are no real populations that ever (even approximately) satisfy these criteria, yet the criteria are required for legitimate application of many population-analytic formulae. In general, however, it is often too difficult to mathematically model the actual dynamics of a real population, so this simplified model is used.

Factors that can cause a population to violate HWE include mutation, gene migration, genetic drift (where the balance of genes in a population changes over time, particularly in small populations), nonrandom mating, population bottlenecks or founder effects, and natural selection. These
commonly occur in real human populations – in particular, the HWE requirement of random mating is violated due to the existence in any geographic region of numerous demographic barriers such as race, religion, language, and physical barriers such as mountain ranges and oceans. Migration rates have usually been low until recent history, but there have been many sudden large-scale migrations corresponding to events in world history. Thus, even an approximate 1-to-1 mapping between geographic populations and genetic populations may not exist. It is important to observe that a non-HWE population is defined in terms not only of the specific combination of genes present, but also the time period that is being considered (since a population changes over time).

Interestingly, it is exactly the differential between HWE and the actual dynamic history of a real population that exposes the intrinsic structure of interrelatedness of a population. Eventually, advances in analysis of these effects will allow for family histories to be reconstructed from descendants’ DNA.

7 Conclusion

We have described the field of molecular genealogy, which is the process of using the combination of genetic and genealogical data to reconstruct the unknown genealogies of living individuals. The relationship of common genetic concepts to molecular genealogy was discussed. Progress has already been made in using genetics for genealogical purposes, particularly with the Y chromosome, which is immediately useful to genealogists because of the correlation of its inheritance mechanism to that of surnames in many societies. Current algorithms for approximate reconstruction of haploid (maternal or paternal) lineage trees were covered, as well as clustering of autosomal DNA to determine population membership. Haplotyping of autosomal and X chromosome loci has been shown as a mechanism to increase the specificity of genetic signatures. Issues of “identity linking” and data accuracy in genealogical data were addressed, in light of the importance of genealogical data to molecular genealogy. Relevant concepts from population genetics were covered. Overall, much progress has been made in developing the tools and concepts that are needed for molecular genealogy, and specific DNA analysis techniques for genealogical research exist today, such as the Y chromosome surname search. However, this field is still in its infancy, and much work still needs to be done to enable genealogists to supplement traditional genealogical research with genetic analysis techniques.

References

Software modeling of the Complement System and its role in Immune Response

Aparna Srikanthswamy, Jugal Kalita, Kaushal Chandrasekar, Karen Newell, AND Patricia Giclas
aparnabs@hotmail.com, kalita@pikespeak.uccs.edu, kchandra@uccs.edu, mnewell@uccs.edu, AND giclasp@njc.org

Abstract

The complement system refers to a series of proteins circulating in the blood that do the work of complementing antibodies in destroying bacteria. These proteins circulate in an inactive form, but in response to the recognition of the molecular components of the pathogenic micro-organism, they become sequentially active, working in a manner wherein the activation and binding of one protein leads to the activation of the next protein in cascade resulting in the lysis of bacterial cells. There are three pathways of complement activation namely classical, lectin and alternative pathways. The initiation of complement cascade in classical pathway is caused by formation of immune complexes on the surface of the target cell. The alternative and lectin pathways are integral participants of non-specific innate immunity and hence do not require any antibody involvement for activation.

Using software, we have developed a model called SIMISYS version 0.4, which implements the activation of the complement system by the three pathways and demonstrates the lysis of bacterial cells through the common membrane attack pathway. It gives the statistics of the number of complement system components involved and the number of bacterial cells destroyed. The software model illustrates other beneficial immune defense functions carried out by the complement system components, namely chemotactic attraction of phagocytes to the infection site, promotion of opsonization, triggering inflammation and removal of immune complexes from the circulation. Complement deficiencies and disease scenarios involving different kinds of bacteria have been simulated using this model. The performance of each pathway of activation is evaluated individually by running the simulation inputting only the required components and making the necessary changes. The software model simulates thousands of biological entities consisting of human cells, bacterial cells and complement system components. A graphical depiction occurs simultaneously on a small panel of the screen as the simulation proceeds. A console output gets displayed indicating the activated pathway, the complement system components involved in it, the bacteria concentration on the terminal. The software written in C++ language is a modification and enhancement of SIMISYS version 0.3 [6], which models and simulates the basic behavior of the human immune system.

1 The Immune system

Any immune response constitutes recognition of the pathogen or other foreign material and the mounting of a reaction against it to eliminate it. An antigen is recognized as foreign when epitopes of that antigen bind to B cells and T cells by means of epitope-specific receptor molecules whose shapes are complementary to that of the epitope. Immune response can be categorized either as innate immune response or adaptive immune response. The innate immune system is a less specific component of the immune system and forms the first line of defense against infection from foreign micro-organism. Pathogen associated molecular patterns can also be recognized by a series of soluble pattern recognition receptors in the blood that function as opsonins and initiate the complement pathways. Some of the cells involved in an innate immune response include neutrophils, macrophages, mast cells, and basophils. Complement proteins and cytokines are some of the molecules involved.
The adaptive immune system is the more specific component of the immune system and is capable of specifically targeting and eliminating a foreign pathogen. It is capable of remembering the pathogen and can build a fight against the intruder preventing it from causing disease later. Adaptive immune response improves by repeated exposure to a given infection and involves antigen presenting cells such as dendritic cells and macrophages, cytotoxic T cells, helper T cells and B cells and production of molecules like antibodies and cytokines [5]. However, the drawback of the specificity of adaptive immune response is that only a few B cells and T cells in the body recognize any one foreign epitope. These few cells then must rapidly multiply in order to produce enough cells to stage a defensive attack against that particular epitope which usually takes several days. This duration of time lets the pathogen to cause considerable harm and hence innate immunity is very essential as the first line of defense [5].

2 Complement System

The complement system is part of the innate immune system and forms the basis of antibody mediated immunity. Apart from defending against bacterial infection, it has other physiologic activities such as bridging innate and adaptive immunity, causing inflammatory response and disposing off immune complexes. Complement was first identified as a heat sensitive principle found in the blood serum that complemented antibodies in destroying bacteria [5]. There are more than 30 complement proteins in the serum. The ways in which the complement system is activated are known as pathways. There are three pathways of activation namely classical, lectin and alternative pathways. The initiation of the complement cascade in classical pathway is caused by the binding of an antibody to an antigen on the surface of the target cell. The alternative and lectin pathways provide non-specific innate immunity which does not require any antibody [5, 13, 15, 16]. All the pathways of activation produce a key enzyme called C3 convertase. C3 convertase cleaves C3 into C3a and C3b [5, 13, 15, 16]. Normally the larger fragment is designated with a $b$ attached to the component name and the smaller is designated with an $a$ attached. All of these three pathways essentially converge at a particular stage in the cascade and produce a common complex called C5 convertase which starts a new set of biochemical reactions ultimately attacking and lysing the target cells. This terminal process is called the membrane attack pathway or the lytic pathway, as it is responsible for lysis [5, 13, 15, 16]. The membrane attack complex forms a large channel through the membrane of the target cell, enabling ions and small molecules to diffuse freely across the membrane [5, 15, 16].

3 Implementation Overview

The software model has been implemented using object oriented C++ language on Red Hat Linux 9.0. It has a multithreaded architecture based on pthreads [10]. The graphics are displayed using Simple Direct Media Layer (SDL) version 1.2 [8]. An XML file has been used to hold different types of complement components, keeping in mind that the system will evolve over time. The Iksemel parser [1] has been used to parse this XML file. The resulting software model is able to simulate an effective defense mechanism by the immune system with the aid of the complement system against a bacterial attack. This model explores the different pathways of the complement system and is able to model disease scenarios. The software model is integrated into SIMISYS version 0.3 [6] and is given the title SIMISYS version 0.4. It includes a graph, depicting the results of the simulation being displayed on a small panel on the screen. A console output gets displayed simultaneously indicating the activated pathway, all the complement system objects involved in it, the bacteria concentration, etc., on the terminal. The model allows the user to be able to zoom in and zoom out using the respective keys on the keyboard. There are three different screens, each representing a different region of the human body. These are the main screen representing the tissue, a second screen representing the activities by the lymphnode object and a third screen displaying the activities carried out by the liver object. The T key on the keyboard enables one to toggle from one screen to another. Most operations carried out in this modeling are string based operations. Pattern recognition through substring recognition and matching of the epitope or signature of each object is the main idea used for movement, opsonization and immune complex clearance mechanisms. The software model includes programs, which implement the complement system as a vital part of the immune system. In this model, a three dimensional
Grid acts as a place holder for the biological entities and molecules. Each grid cell or position can have a pointer to a biological entity like a human cell or a bacterial cell and to a complement system object. Each complement system object references the grid through a grid pointer and the grid keeps track of complement system object present within its cells through a cs pointer. A grid pointer is used by each entity to point to the grid position that it is situated in. Using the grid pointer, a cell can check its neighboring grid positions for other entities, or inquire about the chemicals, and antibodies present within its own position. A back end pointer (bePtr), present in each grid cell points to the entity that is in it currently. The class GridWrapper encapsulates a three dimensional area of grid cells. It contains information about a grid along with the memory allocated to it. The GridWrapper class is helpful while simulating more than one grid area. Every entity and complement system object contains a pointer to a GridWrapper class giving them access to the data in the world to which they belong.

4 Software Specifics

The software modeling of the complement system has been performed by modifying and enhancing SIMISYS version 0.3 [6]. The new version, SIMISYS 0.4 has many new classes providing additional functionality. The proteins and glycoproteins, which constitute the complement system, are mainly synthesized by the liver. Hence a class liver has been created. This class is responsible for creation of components instances and placing them under the GridWrapper representing tissue. A DOM parser is used to parse the XML file which holds the configuration of different types of complement system components to be created along with the count of each such component as type integer. Instances of class components, which are created by the liver object, are placed at different positions randomly in the Grid area representing the tissue. The main functionality of these components instances is to inspect for entity. If an entity of type Bacteria is found, based on its status whether antibody-coated or hasMannoseGroups is set to true. If Bacteria membrane hasLPS, the corresponding pathway namely classical, lectin or alternative and the corresponding initial component namely C1 or MBP or C3b gets activated by setting its respective status to ACTIVE. The components object then appends a string representing the name of current type of component to the cellMembrane attribute of the Bacteria object, which is also a string. This is to indicate the presence of the particular components object and thus aiding in the activation of the next type of components object in cascade. Then it frees itself from the Grid. Now the components object references the Grid through the Bacteria object, so that wherever the Bacteria object moves, the components object moves along with it. In all the cases, only if the MAC-FORMED flag of the components object is set to true, then the status of Bacteria object is set to DEAD. Some basic methods of class components include move(), live(), setStatus(), getStatus(), setType(), getType(), cleave(), releaseAnaphylatoxin(), and die().

In SIMISYS 0.4 some new entities have been introduced to enforce the required functionality. These include the KupfferCell, Erythrocyte, Mastcell and EpithelialCell. All of these entities are created using BasicCell as their super class except for KupfferCell which has been created as a type of Macrophage. This version of SIMISYS still maintains the cellular automata approach of one grid cell can hold one biological cell and one or more identified molecules [6]. Modifications have been made enabling a biological cell or an entity and a complement system object to occupy a single grid cell at any given time. Class erythrocyte has been introduced mainly for the purpose of immune complex clearance [4]. The CR1 receptor is a 4-bit string assigned to each erythrocyte object. It has a value of 1100 which is complement of the C3b receptor value. This CR1 receptor is represented using a character array as shown in Figure 2. If the circulating immune complexes are not cleared, they get deposited in the tissue and cause excessive tissue damage. When an erythrocyte finds a BasicCell object whose status is Antibody-Coated or DEAD or LYSED, it calls the complement() method and tries to find the complement of its CR1 on the cellMembrane of the found BasicCell object.

Figure 1: Character array holding signature of C3b

Figure 2: Character array holding CR1 receptor of Erythrocyte entity
complement of the bit string receptor or signature is found when the 0s of the source signature tallies with the 1s of the destination signature as shown in Figure reftable:crreceptorcomplement. If it finds the complement, the erythrocyte object sets BasicCell objects status to CLEARED, bePtr to NULL, grid pointer of the BasicCell to setOccupied(0) and lastly the gPtr to NULL. This is to make sure the dead Bacteria object does not occupy any Grid place of tissue. It then sets an IC pointe to this BasicCell object, sets its status to CARRIER and move towards the BloodVessel which finds these erythrocyte instances in its immediate vicinity and translocates them to the liver Grid area. The phagocytic Kupffer cells [9] in the liver destroy the immune complexes carried by the erythrocytes.

Because of hydrolysis of C3, it is cleaved to C3b. Some Bacteria instances with hasMannoseGroups flag set to true and many with hasLPS flag set to true are created randomly. As a result both lectin and alternative pathways get activated immediately without any antibodies required. The first reaction to bacterial invasion, in any injury, is a general inflammatory reaction. This can be seen by the number of erythrocyte and Neutrophil instances shooting up to indicate more blood and hence more immune cells around the infected area. This happens after MastCell instances sense a particular concentration of anaphylatoxin and degranulate to produce Histamine which makes the BloodVessel object morePermeable to allow more erythrocyte and neutrophil instances.

Blood vessels in the area of the infection widen to increase the supply of white blood cells that fight infection. The Helper objects get activated and stimulate the BCell objects to produce antibodies. Antibody production and hence the antibody coating of the Bacteria objects activates the classical pathway. All the three pathways are activated. The complement system objects lyses the Bacteria cells directly or chemotactically attract phagocytes which destroys the bacteria cells by opsonization. Due to the combined defense mechanism of the immune system and the complement system the Bacteria entity in this situation is more susceptible and hence its graph starts declining as their number drop.

Lectins are proteins that recognize and bind to specific carbohydrate targets. The lectin pathway is activated by binding of mannose binding protein to mannose residues on the surface of microorganisms [1]. The results in Figure 5 demonstrate a situation where only the lectin pathway components namely MBP, MASp1, MASp2, C4, C2, C3, C5, C6, C7, C8 and C9 are present. Since all the bacterial cells do not display mannose residues on their surface, some Bacteria objects are created with hasMannoseGroups flag set to true. As a result the lectin pathway is not so effective in eliminating the Bacteria but the number of Bacteria objects is kept under control by not letting them to multiply rapidly.

Alternative pathway components are the created in the simulation without creating any other component. Figure 7 illustrates the alternative pathway whereby C3b is being loaded assuming that C3 has been cleaved because of hydrolysis. The result it good as the alternative pathway gets activated through

5 Results

The notations used to label the linear graphs representing different objects are given in Figure 4. Figure 5 illustrates the behavior of the complement system against a simulated bacterial attack through all the pathways. Water is a constituent of the grid.
Classical pathway is activated by antibody-antigen complexes. Figure 8 depicts the results of only classical pathway being activated. This scenario is created with classical pathway components being loaded into the grid and creating maximum number of Bcell and Helper objects specified. The antibodies are loaded into the grid as well and directed against the bacterial cells. This suggests that the probability of all the Bacteria objects in the simulation getting coated with antibody is more. Hence the result is good as there is a huge drop in the number of Bacteria instances.

C3 is a key component in the complement system which generates C3b. It is an abundant serum protein. C3b enhances phagocytosis and enables immune complex clearance. The result in Figure 9 shows a situation where C3 is not created in the simulation. The increase in the count of Bacteria objects is obvious and can be seen in the graph. Patients with C3 deficiency have the most severe clinical manifestations, reflecting the central role of C3 in activation of C5 and formation of the MAC. The majority of patients with C3 deficiency has recurrent bacterial infections and may have immune complex diseases [5, 13, 14].

The graph shown in Figure 10 depicts the secondary complement deficiency in systemic lupus erythematosus [3, 2] disease scenario. The autoantigens from apoptotic epithelialcell objects drive the Macrophage objects to engulf them and present this autoantigenic signature to the Tcell and Bcell objects, which get activated, and start proliferating. Hence autoantibodies are produced. This can be seen by the linear graph of antibodies going up even without the involvement of Bacteria entity. As a result of production of autoantibodies immune complexes are formed with healthy and apoptic epithelialcell objects. Immediately after the immune complexes are formed, the complement system get activated through the classical pathway. Because of excessive consumption of complementsystem objects the linear graph representing complementsystem goes on declining, leading to an impaired clearance of immune complexes. This increase in the concentration of immune complexes can be depicted from the graph.

Late complement system components namely C5, C6, C7, C8 and C9 are required for the formation of membrane attack complex and hence the lysis of target cells. Deficiencies of late complement components [12, 7, 11] that constitute the membrane attack complex result in a significant increase in susceptibility to
neisserial infections. The graph in Figure 11 shows an enormous increase in the number of Bacteria objects when concentration of C5 component is reduced to 0. The count of the Bacteria entity reaches its maximum within no time because of no lysing activity by the complement system objects. Future work will involve workign with biological models of disease for verification and software implementation on a parallel cluster of computers for greater model and simulation realism.

6 Conclusions

This paper has discussed the implementation of a software system called SIMSYS 0.4 that adds a detailed software model of the Complement System to an existing software model of the immune system. The current system models the three complement activation pathways. It is also able to model disease scenarios where the complement system plays significant role.

References


Ivan N. Rich, PhD
HemoGenix, Inc
Colorado Springs, CO, USA

Drug development is a process that can take 10-15 years or more, cost hundreds of thousands of dollars and is one of the most risky endeavors, not only for the company producing the drug, but also for the patients who will receive it. Drug development is really the second stage in the process, the first being drug discovery. It is a very rare occurrence that a single compound is discovered that will turn into a successful drug without testing hundreds or even tens of thousands of different compounds.

Drug development consists of several stages, each of which whittles down the number of compounds until one or perhaps a few show potential promise to be tested in clinical trials. The first stage is screening which may comprise of several rounds of so-called ADME-Tox screening. High-throughput tests are used to determine initial adsorption, distribution, metabolism and elimination of the compound, plus assays that determine potential toxicity. These procedures reduce the number of compounds to about a hundred or perhaps less. The next stage is called lead optimization in which the number is further reduced by about 10 fold. The few remaining compounds then enter preclinical studies in which two or more animal species are used to obtain information that would indicate whether the compound will be efficacious for a specific human disease. The preclinical studies also determine potential toxicity as well as an indication of the starting dose for clinical trials. If all the information is approved by the Food and Drug Administration (FDA), then the compound can proceed to clinical trials of which there are three phases.

During drug development, one of the major goals is to determine or predict the toxic side-effects of a compound and to establish efficacy and safety. Many of the different types of toxicities (hepatotoxicity, cardiotoxicity, gastrointestinal toxicity, nephrotoxicity, neurotoxicity, ototoxicity, pulmonary toxicity, bladder toxicity and endocrine toxicity) can be determined prior to preclinical testing using primarily established cell lines and perhaps even fresh primary cells, for example, hepatocytes. New technology and procedures are being adopted to improve the predictive value of toxicity testing. Hemotoxicity testing, however, has lagged behind and is not given a high priority. There are several reasons for this. First, even though assays have been available since the min-1960s to improve hemotoxicity testing, as discussed in detail below, they suffer from many drawbacks which preclude them from being used in a routine manner. Second, conventional hemotoxicity testing is only performed during preclinical studies and therefore late in the drug development pipeline. Third, conventional hemotoxicity testing involves determining the morphological effects of a compound on the mature blood cell types and blood parameters such as hemoglobin, hematocrit, pH etc. In addition, the pathological appearance of the bone marrow and the spleen is also taken into account. Fourth, these parameters have little, if any, predictive value, because the effects observed in the periphery are the result of actions that have occurred in cells at an earlier point in time and these cells cannot be morphologically identified because they are so few in number. Fifth, if anti-cancer drugs are being developed, these usually
affect the proliferating organs and tissues and it is assumed that the blood-forming system is going to be adversely affected and therefore hemotoxicity, together with gastrointestinal toxicity, is an accepted side-effect of the agent. Finally, the guidelines for hemotoxicity testing, which include these conventional parameters, have not been updated to include the vast amount of biological and physiological knowledge that has accumulated over the decades since the 1940s and 1950s.

To understand how hemotoxicity testing can be used effectively, it is necessary to consider some of the biology involved. The effects of radiation and a large number of cytotoxic drugs have been part of the experimental and clinical research hematology arena for over 60 years. These studies, both in vivo and in vitro, have been used to determine the organizational structure, hierarchy and kinetics of the hematopoietic system as we know it today. Numerous other studies investigating the epithelial cells of the gastrointestinal tract, reproductive organs, epidermal cells of the skin, and the corneal epithelial cells of the eye have shown that, together with the hematopoietic system, a common organizational structure exists for all five continuously proliferating systems. This structure can be divided into 4 main compartments consisting of the functional mature cell compartment, the maturation compartment, the amplification and differentiation compartment and the stem cell compartment. Of prime importance are the amplification and differentiation and stem cell compartments. These compartments contain cells of widely differing proliferative potentials and are therefore affected by extraneous perturbations to differing degrees. However, whereas the first four proliferating systems mentioned above produce just one functional cell type, the hematopoietic system can produce at least 8 different functional cell types (erythrocytes, platelets, neutrophils, basophils, eosinophils, macrophages, T-lymphocytes and B-lymphocytes). All of these cells are produced from a very small pool the hematopoietic stem cells residing in the stem cell compartment of the bone marrow. Stem cells are also present in the peripheral blood and advantage is taken of this fact when mobilized stem cells in peripheral blood are used for hematopoietic stem cell transplantation.

The hematopoietic stem cell compartment consists of a hierarchy of stem cells, from the most primitive and quiescent (in G0 of the cell cycle) to the most mature. An unknown “decision making” process is responsible for maintaining the continuous production of approximately 2 million red blood cells and 200,000 white blood cells every second, and determines whether a stem cell remains in a primitive quiescent state or divides to produce either more stem cells, a process known as self-renewal, or whether it is destined to eventually leave the stem cell pool and become “determined”. For the present discussion, it is important to realize that a mature stem cell has the capacity to divide many times, but this capacity is significantly less than that of a primitive stem cell. Furthermore, the more mature a stem cell is, the greater the chance of it being sensitive to perturbation or compound administration, especially if that compound exhibits an anti-proliferative effect.

By a process known as “stem cell determination”, a mature hematopoietic stem cell enters one or other of the cell lineages leading to a functionally mature blood cell. The immediate progeny of the “determined” stem cell is the progenitor cell which gives rise to the precursor cell. The precursor cells complete differentiation and enter maturation to produce functional mature blood cells.
The proliferating cells that make up the stem cell or amplification and differentiation compartments are the most important targets for hemotoxicity. The potential to use in vitro hemotoxicity began in 1966 when Bradley and Metcalf in Melbourne, Australia and Pluznik and Saks in Rehovot, Israel independently discovered an in vitro population of cells that produced granulocytes and macrophages in culture in 1966, they designated the population colony-forming unit – culture or CFU-C. We now call this biopotent, progenitor cell population the granulocyte-macrophage colony-forming cell or GM-CFC. In 1973, the red blood cell-forming equivalent of the GM-CFC was discovered which was called the burst-forming unit – erythroid or BFU-E. Two years earlier, the in vitro progeny of this erythropoietic progenitor cell population had been discovered and was called the colony-forming unit – erythroid (CFU-E). Indeed, all colony-forming cell populations have been designated in this manner. The erythropoietic precursor population (CFU-E) is equivalent to the first morphologically identifiable erythroblast, the pronormoblast. In vitro progenitor and precursor cell populations have now been discovered for every hematopoietic cell lineage. More importantly however, several in vitro stem cell populations have also been discovered. The most mature of these is the multipotent colony-forming unit – granulocyte, erythroid, megakaryocyte, macrophage or CFC-GEMM stem cell population. As its name implies, this stem cell is capable of producing cells that can populate all hematopoietic cell lineages. A more primitive stem cell called the colony-forming cell – blast or CFC-blast, is considered to be the dividing point between the production of lymphocytes (T- and B-cells) and hematopoietic cells.

Each of the in vitro hematopoietic populations is distinguished by the number and types of growth factors and/or cytokines that are added to the culture, and the incubation time required to produce in vitro colonies of differentiated cells. In general, the more primitive a cell, the more growth factors needed to stimulate it to produce colonies and the longer the incubation time required to produce those colonies. Although recombinant growth factors are now used and serum-free conditions can be employed, the method used to grow different populations has changed little since its inception in 1966. The basic premise for growing hematopoietic cells as colonies in culture is based on the fact that stem and progenitor cells can only be detected by their functional ability to proliferate and divide to produce single colonies of cell progeny. The formation of colonies relies on suspending the cells in a semi-solid matrix that allows the cells to divide, but immobilizes their movement. Several semi-solid media have been used over the years including, agar, plasma clot and water-soluble methyl cellulose. The dividing cells remain stationary to produce aggregates or clusters of cells. Eventually colonies are formed. As the colony grows with time, proliferation gives way to differentiation so that after an optimal incubation period, the colony contains functionally differentiated cells. This technique is called the colony-forming assay and has been used and validated worldwide.

There have been many published reports that have incorporated the colony-forming assay during drug development to determine toxicity. However, they have never been used routinely to examine hemotoxicity. There are several reasons for this. First and foremost, the assay requires manual enumeration and differentiation of colonies. This is a highly subjective, time-consuming operation that requires a great deal of expertise. The procedure lacks standardization and high-throughput capability. Depending on the target tissue and cell population to be detected, the assay can take from 1-3 weeks to perform. From a drug development perspective, they are expensive and only a few companies have the capability of incorporating these assays into their
drug development program. Despite these major drawbacks, the colony-forming assays are the only tests available to detect and measure primitive hematopoietic cells and toxicity to them resulting from drug administration.

It is obvious that the earlier toxicity can be detected in the drug development pipeline, the faster, more efficient and cost-effective the development process can be. For hemotoxicity to play a significant role in this process and not simply be used when the need arises or at the end of the process when a lead compound is being studied pre-clinically or during clinical trials and toxicity is discovered after years of work and large amounts of money have been spent, a new concept had to be devised. This concept involved accepting the drawbacks of the classical colony-forming assay and almost completely redesigning the assay so that the needs of modern drug development were met. More importantly, even though animal target cells could be used, the assay needed to be based on fresh human hematopoietic tissue. The reason for this was obvious. Different species exhibit different toxicities which cannot be directly extrapolated to the human. The goal of drug development is to produce a compound that demonstrates efficacy and safety in the human patient and not in an animal. Therefore, the goal of the new assay was the detection of hemotoxicity in human tissue at the earliest possible stage of drug development and to be able to monitor potential hemotoxicity at any stage in the pipeline, from screening to clinical trials.

The only similarity this new concept has with the classical colony-forming assay is the same ability of primitive hematopoietic cells to be stimulated with different growth factors to proliferate and begin to form colonies in methyl cellulose as the semi-solid immobilizing medium. The new assay, HALO (Hematopoietic / Hemotoxicity Assays via Luminescence Output) is a proliferation rather than a differentiation assay. It is based on the fact that as cell proliferate they increase their intracellular ATP concentration proportionally. After a predetermined incubation time, the cells are lysed to release the intracellular ATP which then reacts with a reagent containing luciferin and luciferase to produce bioluminescence. Since HALO is performed in a 96-well plate, the highly sensitive luminescence readout is measured in a plate luminometer and therefore no manual enumeration is required. The multi-well format allows automation and high-throughput capability. However, there are more important advantages to this platform. First, the assay is standardized and non-subjective. Second, HALO measures proliferative capacity and not differentiation. Results are therefore obtained in a shorter timeframe. Indeed, for human cells, the assay is performed in half the time of the classical colony-forming technique. Third, HALO has multifunctionality. At the present time, 14 different lympho-hematopoietic cell populations can be measured simultaneously. These include 4 stem cell populations and 10 progenitor and precursor cell populations. In addition, all of these populations can be analyzed from 5 species (human, non-human primate, dog, rat and mouse). The multifunctionality of the HALO Platform allows not only measurement of proliferation or inhibition, but also apoptosis/necrosis as a possible mechanism of action of a compound. By comparing the effect of virtually any drug or compound on cells from fresh human peripheral blood or bone marrow with umbilical cord blood, the sensitivity of adult and fetal tissue to a compound can be readily assessed. By comparing compound dose response curves, the IC50 or IC90 values can be ascertained and the compounds ranked in order of their sensitivity to different lympho-hematopoietic populations from different species. The ability to determine the effect of a compound on fresh human hematopoietic tissue not only provides an indication of potential toxicity and efficacy, but also provides a better indication of the starting dose for
clinical trials. Finally, continued monitoring of the patient’s hematopoietic status during clinical trials provides an added safety factor that cannot be underestimated.

No other tissue of the body can be so easily obtained with so little risk and provide such a wealth of beneficial, predictive information as the cells of the blood-forming system. With the introduction of the HALO Platform it is now possible to obtain information on potential hemotoxicity of new drugs at the earliest stages of drug development, thereby reducing risk and providing added safety to the patient.
Two Experiments in Biological Term Annotation using Classification Methods

Sittichai Jiampojamarn, Vlado Kešelj and Nick Cercone
{jiamoja, vlado, nick}@cs.dal.ca

Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada B3H 1W5

Abstract

The number of publications in biological research is growing rapidly. A large part of our research has been involved in making this biological research more accessible. Most of these publications are available online in an unstructured textual format, such as in the PubMed’s MedLine web site. Reading every available article is a time-consuming task; an automatic method of extracting information from them is desirable. The first task in discovering knowledge from these publications is to identify biological terms in articles. This problem is challenging, involving issues such as tackling unknown words, long multi-word terms, and variant author styles to express biological terms.

We present two methods that extend classification algorithms to annotate biological terms. In the first experiment, we exploit special lexical and morphological characteristics of biological terms using different classification algorithms. The C4.5 classification algorithm showed outstanding performance. In the second experiment, we performed C4.5 classification to learn characteristic patterns of biological terms. We obtained 0.76 F-score in extraction of biological terms from the GENIA 3.02 corpus, which includes 2,000 paper abstracts.

Keywords Information extraction, Bioinformatics, Text mining

1. Introduction

Exciting new research into facets of biology had led to a tremendous surge in the amount of biological literature published. Most of these publications are available online such as in the PubMed’s MedLine database [1]. Providing software support for knowledge discovery from the available publications in the biomedical area is a challenging endeavor. Knowledge discovery from online biological publications is difficult because the articles are typically unstructured text. The first step in making these articles more accessible is to annotate the biomedical terms, pointing out the interesting words in the documents, which are the most salient content-bearing words. After this step, one can proceed with detection of relations between the terms, pathways discovery from the literature. However, reading all articles and annotating manually by human experts becomes an infeasible task due to the large size of data. Automatically annotating the biological terms in these unstructured articles becomes a crucial task. The following is an example shown input text [2] and output text where biological terms are annotated in documents:

Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation.

Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation.

Machine-learning methods are used to automatically annotate biological terms, such as gene names, proteins, organs, diseases, etc.—i.e., a list of terms that is specified by domain experts. This task is similar to the named-entity recognition task in the Message Understanding Conference (MUC). Difficulties in identifying these terms are included on unseen terms, long multi-word terms, and various writing styles by different authors.

Biological terms can be single-word terms (e.g., Adenovirus, E1A, tublin, GATA-1) or multi-word terms (e.g., mouse interleukin-1 receptor alpha gene, large granular lymphocytes). Many multi-word terms can be written in different styles depending on the author’s preferences. For example, some authors prefer abbreviated short terms, and

1 The proceedings of the seven MUC conferences were published by the Morgan Kaufmann Publishers in 1990’s. The Web site is http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.
other prefer the full long form; e.g., “large granular lymphocytes” can be re-written as “LGL,” or “UDG” as “uracil-DNA glycosylase”.

To solve this problem, we used classification algorithms to identify biological terms by learning regular characteristics of terms rather than manual setting patterns or looking-up any biological dictionary. Our approach is based on relaxing fixed rules and generalizing the system to extract biological terms in different styles of research publications. We take advantage of special character-based biological term characteristics, including uppercase letters, digits, and symbols, as well as the biological concept words such as “gene,” “cell,” “protein”, and so on. Part-of-speech tagger [3] is also applied in automatic classification methods to learn the patterns of these special characteristics in biological terms.

The paper is organized as follows. In the next section we introduce some related work. In Section 3, two experiments in the problem of biological term annotation are described, and the results are presented and discussed. The conclusion and future work are presented in Section 4.

2 Related Work

Text mining in biology has been interested in many research groups and conferences, including the TREC genomics track [4] which has a goal to provide information support tools for genomic domain. A biological term annotation task plays an important role in measuring biologically meaningful shown in BioCreAtIVE [5]. Recently there have been published papers on the problem of biological term annotation and many of them made use of the GENIA corpus [2]. The current methodology in biological terms extraction task can be divided in two main approaches, rule-based methods and learning methods.

K. Fukuda et al. [6] focused on identifying protein names based on traditional rule-based methods in which the rules are manually set by observing some special characteristics in protein names such as uppercase letters, lowercase letters, digits, special symbols and characteristic words such as kinase, receptor, protein, etc. Then they combine those “core terms” together as they describe that most protein names are compound multi-word terms. They obtain 0.94 precision and 0.98 recall on 50 research abstracts retrieved from MedLine.

N. Collier et al. [7] used HMM (Hidden Markov Model) to extract protein and DNA names in a small corpus of 100 MedLine abstracts. They defined the word features based on characteristics of known terms in training data set such as digit numbers, uppercase letters, lowercase letters, Greek letters, combination of uppercase, lowercase letters and digits, and special symbols. They achieved 0.73 F-score on a cross-validation test.

L. Venkata Subramaniam et al. [8] proposed Bioannotator, a biological terms annotation system which combines both a rule-based and a dictionary-based engine. They used a shallow parser to identify noun phrases in the documents, then each noun phrase is labeled whether it is biological term or not based on three dictionaries: Unified Medical Language Systems (UMLS) [9], LocusLink [10], and GeneAlias, as well as the rule engine. The rule engine is a set of regular expressions to recognize a word which contains uppercase letters, digits, special symbols, Greek letters, and characteristic words in biological terms such as amylase, cell, gene, amino, etc. The system achieved 0.64 F-score in exact matching and 0.90 F-score in partial matching with the answers in GENIA 1.1 corpus [2], which contains 670 research abstracts. This result showed that the boundaries of biological terms are difficult to detect. Although most of biological terms are nouns, these terms are usually just parts of proper noun phrases.

Our approach relies on the fact that biological terms usually have some special character-based characteristics, including uppercase letters, digits, and symbols, as well as the biological concept words such as “gene,” “cell,” “protein,” etc. Instead of manually setting rules, we use classification methods to learn the patterns of these special characteristics in biological terms.

3 Methods, experiments, results

We used existing classification methods to identify the terms automatically rather than to set up rules manually to capture some specific characteristics of biological terms. A classifier extracts boundaries of biological terms, starting and ending positions, from the plain text in documents. We used word n-grams to chunk out each sentence and extract feature attributes for training classifiers. The feature attributes are general characteristics of each word in a n-gram. However, this is still a question how to define the significant feature attributes, which capture distinction between biological terms from normal words. We set up two experiments to investigate the proper way to use the classification scheme for biological terms extraction.

3.1 First experiment

It has been shown in [6, 7, 8] that biological terms have some regular characteristics expressed in appearance of the uppercase letters with special symbols or digit numbers. Many terms contain Greek letters, expressed as English words, or have prefixes or suffixes of biological concepts (e.g., ase, cyt). The multi-word terms usually contain characteristic words of the biological domain such as cell(s), protein(s), amino, and gene(s). Therefore, we first set feature attributes to be:
numbers of uppercase letters.

- numbers of digits.

- numbers of symbols.

- numbers of Greek letters shown in table 1.

- indicator of whether a word has a specific biological prefix or suffix concept.

- part-of-speech tag information tagged by bi-gram hidden Markov model POS tagger [3].

<table>
<thead>
<tr>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeta</td>
<td>Eta</td>
<td>Theta</td>
<td>Iota</td>
<td>Kappa</td>
</tr>
<tr>
<td>Lambda</td>
<td>Mu</td>
<td>Nu</td>
<td>Xi</td>
<td>Omicron</td>
</tr>
<tr>
<td>Pi</td>
<td>Rho</td>
<td>Sigma</td>
<td>Tau</td>
<td>Upsilon</td>
</tr>
<tr>
<td>Phi</td>
<td>Chi</td>
<td>Psi</td>
<td>Omega</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Greek Letters

We use a sliding window of $n$ words, i.e., a word $n$-gram model, that starts from the beginning of each sentence shown in table 2 and extracts feature attributes described above for each word. Additionally, with each word position we associate a class attribute that labels the position as a starting or ending position of a biological term, or as an undistinguished position. Classifiers are trained and evaluated on the task of detecting starting and ending positions of a term. This position information is further used in annotating biological terms.

**Input sentence:**

The CD4 coreceptor interacts with non-polymorphic regions of major histocompatibility complex class II molecules on antigen-presenting cells and contributes to T cell activation.

**Word n-gram:**

The
The CD4
The CD4 coreceptor
CD4 coreceptor interacts
coreceptor interacts with

T cell activation.
cell activation.
activation.

Table 2: word n-gram example, where $n = 3$

We trained and tested our method on the GENIA 1.1 corpus [2], which contains 670 research abstracts with biological terms annotated by human experts. We used classification algorithms in the WEKA 3.4 machine learning tool [11] to perform our experiments. The result of extracting the starting positions in different classification methods are shown in table 3. Precision, Recall and F-score are common standards to evaluate the performance of a classifier defined as in equations 1, 2, and 3, where TP (true positive) is the number of terms which are correctly classified to the class, FP (false positive) is the number of terms which are correctly unclassified to the class, and FN (false negative) is the number of terms which are incorrectly unclassified to the class.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{2}
\]

\[
F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}
\]

As shown in table 3, we obtain the best performance from the C4.5 classification algorithm achieving the F-score of 0.64. In the same way, we obtained 0.72 F-score for the ending position accuracy. We combined both information from the classifier to extract biological terms based on starting and ending positions. The resulting performance decreased in accuracy to be 0.52 F-score for exact matching against the answers in the corpus shown in table 4. We explain this effect as an accumulated error from the starting-position and ending-position classification. Either an error in starting or ending position will lead us to a wrong term in the sense of exact matching. If the errors in determining the starting and ending positions were completely independent, we would have a probability of $0.71 \cdot 0.76 = 0.54$ of getting an exact match. The exact matching precision 0.65 is somewhat higher, which implies that detection of starting and ending positions are somewhat correlated. With respect to recall, the independence assumption implies an exact matching recall of $0.58 \cdot 0.69 = 0.40$, which is very close to actual 0.43. This experimental result justifies our conclusion that the decreased performance on exact matching is largely due to independent treatment of starting and ending position detection.

### 3.2 Second experiment

In section 3.1, the experiment revealed us some problems in automatically detecting boundaries in multi-word biological terms. Tracking only starting and ending position of each term does not provide us with enough information to annotate terms in overall, dealing with both long words and single words at the same time.
In the second experiment, we added more target classes for the classifier. Unlike earlier, we classify each instance in five classes as starting, middle, ending, single, and non-relevant. The starting and ending classes indicate the beginning and ending positions of biological terms while the middle class indicates the words between starting and ending positions in a multi-word biological term. The single class means that those biological terms are only one word long. Finally, the other words, which are not biological terms, will be classified as the non-relevant class.

From the first experiment we know that biological terms are correlated with the occurrences of uppercase letters, digits, and special symbols. These features show the difference between general words and biological words as the higher number of uppercase and symbol letters. However, they do not cover the case where short biological words have a small number of uppercase letters or symbols. The words with uppercase letters in the middle have higher possibility to be biology related than uppercase letters at the beginning, which can be confused at the beginning of sentences and in proper nouns. From our observation, the number is not as important as pattern of character features formed in a word.

We further refined the sensitivity of the classification algorithms by extracting word feature patterns, including Greek letters, Leading words, Uppercase, Lowercase, Digits, Hyphen, Plus, Slash, OpenParen, CloseParen, OpenSquare, CloseSquare, Percent, and other symbols. The word feature patterns provide the capability to annotate unseen biological words which have the same pattern even they are not in the training set. To reduce dimensionality, we limit the number of features recorded for each word in an n-gram window to \( m \). For example for \( m = 4 \), the term “CD28-mediated” generates the pattern as Uppercase, Digits, Hyphen, and Lowercase. In case a word has more features than the value \( m \), only the first \( m \) features are extracted and the rest of the features are ignored. In another case, if a word has less features than the value \( m \), the dummy feature “none” is appended up to the \( m \) value. In this case, the number of feature attributes for each instance in the classifier is \( m \times n \), where \( m \) is the value of word features we extract from \( n \)-word n-grams.

We used the GENIA 3.02 corpus which contains 2,000 paper abstracts. As the result in table 5, the classifier identified single biological terms with higher accuracy than multi-word terms indicated by lower accuracy on starting, middle and ending classes. Hence, there is not much difference or unique characteristics among starting words, ending words, and the words in the middle. However, when we combined all classes above as one class labeled “relevant,” the classifier provided a higher accuracy on making the decision whether the word is the biological term or a normal word with precision 0.90, recall 0.65 and F-score 0.76.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaiveBayes</td>
<td>0.54</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>C4.5</td>
<td>0.71</td>
<td>0.58</td>
<td>0.64</td>
</tr>
<tr>
<td>AdaBoostM1</td>
<td>0.67</td>
<td>0.25</td>
<td>0.37</td>
</tr>
<tr>
<td>LogisRegess</td>
<td>0.66</td>
<td>0.45</td>
<td>0.53</td>
</tr>
<tr>
<td>SMO</td>
<td>0.69</td>
<td>0.34</td>
<td>0.46</td>
</tr>
<tr>
<td>IB1</td>
<td>0.69</td>
<td>0.58</td>
<td>0.63</td>
</tr>
<tr>
<td>Rule Part</td>
<td>0.72</td>
<td>0.54</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3: Classifications performance

<table>
<thead>
<tr>
<th>Annotate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting</td>
<td>0.71</td>
<td>0.58</td>
<td>0.64</td>
</tr>
<tr>
<td>Ending</td>
<td>0.76</td>
<td>0.69</td>
<td>0.72</td>
</tr>
<tr>
<td>Exact match</td>
<td>0.65</td>
<td>0.43</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 4: Annotation performance

<table>
<thead>
<tr>
<th>Target Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting</td>
<td>0.58</td>
<td>0.44</td>
<td>0.50</td>
</tr>
<tr>
<td>Middle</td>
<td>0.68</td>
<td>0.26</td>
<td>0.38</td>
</tr>
<tr>
<td>Ending</td>
<td>0.69</td>
<td>0.49</td>
<td>0.57</td>
</tr>
<tr>
<td>Single</td>
<td>0.60</td>
<td>0.61</td>
<td>0.60</td>
</tr>
<tr>
<td>Relevant</td>
<td>0.90</td>
<td>0.65</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 5: Word tag positions performance

4 Conclusions

In this paper, we presented two experiments for biological terms annotation using classification methods. The classifiers captured the regular characteristics of biological terms from training data, then they were used to detect whether the terms are biology related or not. We used different classification algorithms. The experiments demonstrated that the C4.5 algorithm is the suitable classification method for annotation of biological terms. We used general word features such as uppercase and lowercase letters, digits, special symbols, as feature attributes for the classifier to learn biology-related term patterns. These features are general and can be adapted for other domains. We got 0.76 F-score on distinction between biological and normal terms. However, extracting exact multi-word terms remains to be improved in future.

Our results are comparable with dictionaries and rule-based systems [8] while we reduced the manual effort in creating rules and patterns and tested the results on a larger corpus. However, the selection of feature attributes used in classification need to be improved for a better performance, especially in exact annotation of the boundaries of multi-word biological terms.

In future work, we plan to enrich feature attributes with
new general features relevant to this problem as well as consider rough set theory [12] helping in feature selection for classification [13]. The problem of defining boundaries of each term can be addressed by considering syntactic analysis.

References


Unsupervised Evolution of Probabilistic Neural Networks for Clustering Gene Expression Data

Daron Anderson
University of Colorado at Colorado Springs
danders4@uccs.edu

Several techniques have been applied to micro-array data in attempts to provide an automated method of preprocessing data so that scientists studying the data can analyze and interpret it efficiently. Very few techniques employ neural networks or genetic algorithms, so a combination of both is applied here to test the ability of these techniques as a tool for scientists studying gene expression data.

Most techniques require either a subject-matter expert’s knowledge to implement and to interpret the results, or need to be retrained for each new data set. In this experiment, I selected a neural network architecture and training algorithm that has several traits: (1) the clustering algorithm does not require a-priori knowledge of the number of the clusters, (2) the clustering criteria could be read directly from the structure of the network, (3) the resulting neural network has the potential to be applied to other related data sets without being retrained.

The neural network architecture used in this experiment is a variation of the traditional Probabilistic Neural Network (PNN) [1]. A PNN is a three-layer network that uses probabilities to classify input data. Unlike most neural networks, a PNN is not trained. Instead, a set of ‘training patterns’ which are normally existing data elements is used as the weights of the pattern (middle) layer. The Input and Pattern layers of the PNN are fully connected. Each neuron in the pattern layer sums up the squared Euclidean distance of the input pattern from the training pattern for that neuron, and then applies a Gaussian probability distribution to the sum.

The pattern and summation (output) layers are not fully connected. Rather, all of the neurons associated with training patterns that belong to some class A are connected only to a single neuron in the summation layer that represents the probability that the input pattern belongs to class A. To compute the probability that some input pattern belongs to class A, the implementer provides the a-priori probability that some pattern belongs to class A, as well as a misclassification loss that indicates how costly to the end user a misclassification.

The traditional PNN requires a lot of information that is known by the implementer for it to do a good job, and in the world of bioinformatics, the information required by a PNN is not necessarily available. Therefore, I propose a modified PNN topology that is capable of being trained without supervision. This modified topology is intended to better fit a clustering problem. Also, since this architecture is based on statistical measurements, it can be interpreted after the network has been trained.

Since no training patterns are expected to be available, the pattern layer and summation layer are merged into a single output layer. Each neuron in the output layer represents a cluster of data. The weights associated with a neuron in the output layer are actually the centroid and the standard deviation of the cluster. The misclassification loss and a-priori probability are not used, since it is only meaningful if supplied by a subject-matter expert. Since each output neuron represents a single cluster, this topology allows the number of cluster to be evolved, rather than being pre-determined.

A population of modified PNNs is evolved using a modified genetic algorithm. This algorithm uses classic selection and replacement strategies as well as specialized crossover and mutation operators. Specialized operators for inserting and deleting clusters were also employed. The crossover, mutation, cluster insertion and cluster deletion operators each have an associated probability that determines if they are applied during the current network. A special restriction was enforced so that the cluster deletion operator could not be applied to a network with a single cluster.

The selection strategy used for this algorithm was tournament selection of size 4. A group of 4 networks were selected from the population at with equal probability, and the two most-fit networks were selected as the parents. The replacement strategy is local elite replacement. This enforces that the 2 most-fit networks in a breeding event survive to the next generation. For example, if both parents of a breeding event are the most-fit networks, then the neither of the children survives. This ensures that the overall fitness of the population will never be reduced.

The fitness of a modified PNN is computed from a set of input patterns. Each PNN in the population is presented with each input pattern. The cluster predicted for the input pattern, and the Euclidean distance of the input pattern from the centroid of the predicted cluster is computed. The distances for all input patterns are
summed up and the total is adjusted by two penalty values, to become the fitness of the individual PNN. The network with the smallest sum is considered the most fit, which minimizes the distances of input pattern to the centroid of its cluster.

As networks evolve, clusters may be inserted into a network that do not ‘predict’ the cluster of any input data element, or a cluster may become useless as it ‘looses’ input patterns to another cluster in the PNN. Since clusters that do not predict the cluster of any input value are misleading to the scientist who would use this network, a useless cluster penalty is applied to the fitness of the PNN for each useless cluster.

If the genetic algorithm is run for enough generations, it is possible that clusters will be added to the network that fit perfectly to a single input pattern, allowing one cluster for each input pattern to exist in the network. To avoid this, a cluster penalty is applied for each cluster that exists in the network to help prevent too many ‘small’ clusters from being added to the network.

To measure how effective the clustering of the PNNs is, I used the standard Precision and Recall measurements. To calculate the precision and recall of the clusters in the PNNs, I had to make a simplifying assumption. Since the data showed a tendency to not fall into exactly two clusters, I assume that the cluster in a PNN represents class X if the majority of the elements that are assigned to that cluster are of class X. Therefore, several clusters in a PNN might belong to class X. This simplification was done only to allow some meaningful numeric values to be assigned to each cluster. A PNN that actually did a good job of clustering data would still show high values for precision and recall of both of the test classes.

To verify that the genetic algorithm could successfully cluster data, a simple synthetic data set was supplied to the algorithm and evolved. The algorithm scored 100% precision and recall with the data set. Further, the predicted centroids of the 4 clusters were extremely close to the true centroids, indicating that the clustering did find the actual clusters of data, not just the grouping of data points. This performance was much better than the original K-Means clustering, which was only able to find all four clusters in 50% of the runs.

The initial fabricated test data showed much promise in how quickly it converged to a solution, and how well that solution performed on the data set. An open issue with the initial success that was not addressed by the training algorithm is how to compare the value of one standard deviation for a cluster to another standard deviation. For the purposes of this experiment it was assumed that the performance of the evolved PNNs is not affected by this omission.

Golub et al.[2] used data sets containing expression data collected from samples correlated with two types of leukemia: AML and ALL. Golub used a measure of statistical similarity to select the 25 expression sequences that were most closely correlated with the AML, and 25 for ALL. Each sample was normalized into a range from −3 to 3. The mean of the raw data was set to the value 0 in the range, and the standard deviation was measured as the value 1. Therefore, the data was normalized into 3 standard deviations from the mean value before any clustering was attempted.

The genetic algorithm was applied to the Golub data set, with marginal success. First, the initial number of clusters was kept at 1 while the population size, probabilities for mutation, etc., and penalties were varied. This seemed like a reasonable start, since the test data performed very well under these conditions, and managed to find all four clusters.

The networks produced for each set of parameters all had populations that were entirely single-cluster PNNs. All runs of the algorithm converged to about the same fitness value, so it was questionable whether the algorithm did not perform well, or if the data actually belonged to a single cluster.

To test this, a second run was performed with a different set of parameters for evolution. Only the initial number of clusters was modified between runs. The other parameters were selected to emphasize a high level of searching the solution space by setting the mutation probability to 1 and lowering the cluster penalties.

The precision and recall on the training data was marginal. The ALL data tended to have better precision and recall than the AML data. The ALL data had a best precision/recall value pair of 75%/57%, where the ALL data had 50%/62%. These values were obtained when evolved with 3 initial clusters. The average number of clusters at the end of training was approximately 3, regardless of the initial number of clusters, indicating that some natural number of clusters for the data may be 3, instead of the expected 2.

These results are not expected to be optimal since the one-dimensional Gaussian distributions used in the clusters form a ‘hypersphere’ distribution around each centroid. Based on the results, more experimentation with the distribution is required before this algorithm can be applied successfully.

References


In Silica Characterization of Rad26, a Fission Yeast DNA Damage Checkpoint Protein

Kristyn Baker, Jugal Kalita and Tom Wolkow
University of Colorado at Colorado Springs
1420 Austin Bluffs Pkwy
Colorado Springs, CO USA 80918
kristybaker@usa.net, kalita@uccs.edu, twolkow@uccs.edu

1. Introduction

Eukaryotic cells use an evolutionarily conserved pathway to find DNA damage and ensure that it is repaired prior to cell division. Called the DNA damage checkpoint, this pathway surveys the integrity of genomic DNA. Upon sensing damage, it halts mitosis and simultaneously influences the activities of DNA repair enzymes. Cells lacking a functional DNA damage checkpoint initiate mitosis when DNA damage is present and thereby allow damaged DNA to segregate to daughter cells. The checkpoint is critical to the proper development of humans, as mutations in checkpoint genes predispose individuals to numerous cancers.

This study will examine how the checkpoint monitors genomic DNA for damage. Recent findings suggest that the evolutionarily conserved checkpoint protein Rad26 orchestrates this function in fission yeast (Schizosaccharomyces pombe). Upon discovering damage, Rad26 seems to alert downstream checkpoint proteins that initiate cell cycle arrest and DNA repair activities. It is not known how Rad26 performs these important checkpoint functions, mainly because the regulatory and biochemical properties of this protein are not known.

We are using bioinformatics methodologies to help elucidate the regulatory and biochemical properties of Rad26. We predict that these in silico techniques will allow us to identify evolutionarily conserved domains of Rad26 that will elucidate these properties. We have attempted to identify these conserved domains by

♦ Visual inspection of the Rad26 sequence
♦ Comparison of rad26 and its homologs with the BCM Search Launcher multiple sequence alignment software
♦ Comparing the Rad26 protein sequences to other sequences in protein databases with the Basic Local Alignment Search Tool (BLAST)
♦ Scanning Rad26 for well characterized domains using the BCM Search Launcher Coils software.

Molecular genetic techniques, including in vitro mutagenesis, will be used to mutate domains and test if they are essential for Rad26 function. In the future, we plan to characterize the biochemical properties conferred by Rad26 domains and ultimately determine how this protein relays the status of DNA to downstream checkpoint processes.

2. Results

2.1 Inspection of the Rad26 sequence revealed a possible regulatory domain

Visual inspection of Rad26 identified a possible nuclear localization sequence (NLS) [1]. NLSs are commonly regulated by phosphorylation. In Rad26, the site near the potential NLS resembles the conserved consensus phosphorylation site of Cdc2, a highly conserved cyclin dependent kinase required for cell cycle transitions.

♦ NSL Consensus site: P-KKK-K
♦ Cdc2 Consensus site: S/T-P-X-K/R (the S/T is phosphorylated by Cdc2)
♦ Putative Rad26 NLS: SPSKKKRK

We are currently mutating this Rad26 sequence to APSKKKRK to test if this site regulates the nuclear localization of Rad26.

The Prospero v1.3 software tool performs self-alignment that can indicate conserved domains. Thirty or more consecutive residues are considered significant [2]. Rad26, the homologs, and the chaperone protein did not
contain any significant repeats when processed by Prospero (data not shown).

2.2 Multiple sequence alignments revealed a possible domain in Rad26

Multiple sequence alignment comparisons were made among the proteins encoded by rad26$^+$ and its homologs; Saccharomyces cerevisiae ldc1$^+$, human ATRIP and uvsD$^+$ of Aspergillus nidulans. A chaperone was included as a negative control to determine if the results were significant. Table 1 lists the protein sources.

Table 1. The Entrez protein sources

<table>
<thead>
<tr>
<th>Protein</th>
<th>Organism</th>
<th>Locus</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rad26</td>
<td>S pombe</td>
<td>P36632</td>
<td>6174941</td>
</tr>
<tr>
<td>Lcd1p</td>
<td>S cerevisiae</td>
<td>NP 010787</td>
<td>630707</td>
</tr>
<tr>
<td>ATRIP</td>
<td>H sapiens</td>
<td>Q8WXE1</td>
<td>48428109</td>
</tr>
<tr>
<td>UvsD</td>
<td>A nidulans</td>
<td>AF180367</td>
<td>5853268</td>
</tr>
<tr>
<td>chaperone</td>
<td>S pombe</td>
<td>CAB38512</td>
<td>4481954</td>
</tr>
</tbody>
</table>

We found that Rad26 was most similar to UvsD followed by Lcd1p and ATRIP. As expected, the chaperone protein shared no significant similarities with Rad26. Refer to Table 2 for details. The raw score [3] is the alignment score calculated as the sum of substitution and gap scores. The probability value is the probability of an alignment occurring with the score in question or better. The most significant $P$ values are those close to zero.

Table 2. Alignment of Rad26 and homologs

<table>
<thead>
<tr>
<th>Protein</th>
<th>Raw Score</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATRIP</td>
<td>30.3</td>
<td>0.21</td>
</tr>
<tr>
<td>Lcd1p</td>
<td>37.6</td>
<td>0.00079</td>
</tr>
<tr>
<td>UvsD</td>
<td>51.2</td>
<td>7.3 x 10$^{-7}$</td>
</tr>
<tr>
<td>chaperone</td>
<td>No significant similarity</td>
<td></td>
</tr>
</tbody>
</table>

The multiple sequence alignment tool was used as designed by submitting the proteins to be aligned into the tool. We observed that these proteins are highly divergent and at most contain 36% identity with the average being 25%. However, we did find a putative conserved block of residues (Table 3). The black background highlights the identical residues while the grey highlights those that are similar.

Table 3. A putative, conserved domain in Rad26

<table>
<thead>
<tr>
<th>Protein</th>
<th>Start</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rad26</td>
<td>88</td>
<td>REG2N-ALF2HALINPR</td>
</tr>
<tr>
<td>UvsD</td>
<td>178</td>
<td>REG2N-ALF2HALINPR</td>
</tr>
<tr>
<td>ATRIP</td>
<td>195</td>
<td>REG2N-ALF2HALINPR</td>
</tr>
<tr>
<td>Lcd1p</td>
<td>128</td>
<td>L009RALK9REPR</td>
</tr>
</tbody>
</table>

We are currently testing if these residues are required for Rad26 function by changing R to A and testing if the resulting protein disrupts the fission yeast DNA damage checkpoint.

2.3 BLAST results identified similarities among Rad26 and PARP family members

BLAST comparisons of Rad26 with standard protein databases identified similarities with several poly(ADP-ribose) polymerase or PARP proteins (Tables 4 and 5). These alignments are intriguing because PARP enzymes can detect DNA damage.

Table 4. PARP domains from the NCBI CDD

<table>
<thead>
<tr>
<th>Pfam</th>
<th>Type</th>
<th>Function</th>
<th>PSSM ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>pfam00644</td>
<td>PARP</td>
<td>catalytic</td>
<td>1202</td>
</tr>
<tr>
<td>pfam00645</td>
<td>PARP</td>
<td>zinc fingers</td>
<td>6182</td>
</tr>
<tr>
<td>pfam02877</td>
<td>PARP</td>
<td>regulatory</td>
<td>3371</td>
</tr>
</tbody>
</table>

Table 5. Alignment of Rad26 with PARP pfams

<table>
<thead>
<tr>
<th>Protein</th>
<th>Start</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>pfam00644</td>
<td>12</td>
<td>SEEPBRIR</td>
</tr>
<tr>
<td>Rad26</td>
<td>89</td>
<td>SEEPBRIR</td>
</tr>
<tr>
<td>ATRIP</td>
<td>339</td>
<td>SEEPBRIR</td>
</tr>
<tr>
<td>Lcd1p</td>
<td>72</td>
<td>SEEPBRIR</td>
</tr>
<tr>
<td>UvsD</td>
<td>215</td>
<td>REELAR</td>
</tr>
<tr>
<td>pfam00644</td>
<td>190</td>
<td>LNEWITLH</td>
</tr>
<tr>
<td>Rad26</td>
<td>472</td>
<td>LNEWITLH</td>
</tr>
<tr>
<td>ATRIP</td>
<td>695</td>
<td>HRQWTVR</td>
</tr>
<tr>
<td>Lcd1p</td>
<td>686</td>
<td>TTRPIPLH</td>
</tr>
<tr>
<td>UvsD</td>
<td>691</td>
<td>LKLHPG</td>
</tr>
<tr>
<td>pfam00645</td>
<td>2</td>
<td>SEYARIK</td>
</tr>
<tr>
<td>Rad26</td>
<td>134</td>
<td>SEYARIK</td>
</tr>
<tr>
<td>ATRIP</td>
<td>351</td>
<td>SEYARIK</td>
</tr>
<tr>
<td>Lcd1p</td>
<td>86</td>
<td>NRESPK</td>
</tr>
<tr>
<td>UvsD</td>
<td>149</td>
<td>ILPSQ</td>
</tr>
<tr>
<td>pfam02877</td>
<td>Not much similarity</td>
<td></td>
</tr>
</tbody>
</table>

The LNEWITLH segment will be mutated because W’s are known to have an affinity for DNA. The mutated segment will be LNEAITALH.

2.4 Rad26 may have a coiled-coil structure

The NCBI Entrez Protein and Conserved Domain databases list known functionality for proteins and families. This information was referenced to predict what functionality the Rad26 significant sequences may have. Unfortunately, other than identifying coils and helices, no additional information was uncovered.
The BCM Search Launcher Coils tool predicts coiled-coil secondary protein structures. Rad26 and homologs all showed a coiled-coil in approximately the same N-terminal region. In contrast, the chaperone protein had several coils staggered fairly evenly throughout the protein. For Rad26 the coil is projected to start around residue 98 and end around 165.

Coiled-coil structures are composed of seven residue heptad repeats. The seven residues are designated by abcdef where a and d are hydrophobic residues [4]. EIEYAKTK was revealed by comparison with pfam00645 and is also part of the putative Rad26 coiled-coil. In this segment the T is in the f position and may mediate Rad26 interactions with other proteins. So, the segment will be changed to EIEYAKAK and then tested for functionality in fission yeast.

2.5. Conclusion

Based on our in silico findings we intend to create the four Rad26 mutations listed below. We will then test if these residues are important for Rad26 function in vitro.

- SPSKKKRK → APSKKKRK
- SGNAILR → SGENAILA
- LNEWITLH → LNEATLHL
- EIEYAKTK → EIEYAKAK

3. In vitro component

The in vitro mutagenesis portions of this exercise are in progress.

5. References


Rapid and Efficient Dissemination of Biological Data with Interactive Keys: The Southern Rocky Mountain Interactive Flora

Jeffrey Wayne Brasher & Neil Snow
Department of Biological Sciences
University of Northern Colorado
Greeley, CO 80639
jeff.brasher@sbcglobal.net, Neil.Snow@unco.edu

Abstract

Writing a traditional Flora for the identification of local plants can take a decade or more. In contrast, interactive keys (IKs) can be produced and updated rapidly, representing an underutilized technology in bioinformatics. IKs accelerate specimen identification and link information at all levels, from ecosystems to molecules. Online IKs democratize science, giving less developed countries access to information globally, and providing local control over data production and distribution. Our Southern Rocky Mountain Interactive Flora uses Lucid. This region will be the first to have an interactive plant key of this size (> 3200 species) online. Conceptual challenges must be overcome. One is the terminology used for plant characters and character states. Four phases of the IK will be released successively, allowing "peer review". Phase 1 is a key to families. Phase 2, genera, will be released soon. Phase 3, species, will follow. Phase 4 will link text, photographs, and a federation of specimen databases.

1. Introduction

Given the current biodiversity crisis, it is the opportunity and responsibility of the scientific community to take full advantage of the rapidly developing possibilities of bioinformatics to expedite and revitalize descriptive systematics, the field of biology that provides basic data about biological organisms, enabling society to understand and conserve the wonders of nature [8, 9]. The traditional process of writing scholarly, in-depth Floras and Faunas for the identification of biodiversity is a slow, often decades-long endeavor. Subsequent editions typically appear only after many years, if ever. In contrast, computerized interactive keys (IKs) can be produced relatively rapidly, and represent an important but underutilized new tool in bioinformatics. Their primary benefits are the ability to rapidly expedite the production of floras and faunas and accelerate the rate at which unknown specimens can be properly identified. They can also be easily updated given their electronic format. We expect IKs will become increasingly field-portable on handheld computers and cell phones.

Various "big science" bioinformatics-related initiatives such as GBIF (Global Biodiversity Information Facility) [4, 9], LINNE (Legacy Infrastructure Network for Natural Environments) [5, 9], and NEON (National Ecological Observatory Network) [5] are underway. IKs will facilitate and complement these efforts by linking biological information infrastructure ranging from GIS ecosystem data and specimen label databases, to behavioral, physiological, and genomic data spanning the tree of life.

IKs made available online have a democratizing effect in science, giving workers in less developed countries web access to information globally, and providing local control over data production and distribution.

Our Southern Rocky Mountain Interactive Flora (SRMIF) ("sir-mif") is leading this area of bioinformatics in the USA. The Southern Rockies will be the first region to have an interactive key to plants of this magnitude (> 3200 species) [6] available online. SRMIF is to be a free, full-service vascular plant flora, treating all of Colorado, northcentral New Mexico, and southeastern Wyoming.

2. Materials and Methods

We use Lucid Professional software [2] given its many troubleshooting devices that help the user during the identification process, its ease of use for IK authors, its currency, and its prospects for future development. We judge it the best software available for interactive keys. We now use Lucid Builder 2.1 and Lucid Player 2.2, but the full application version of Lucid 3, to be
released later in 2004 promises to be even more powerful and user-friendly for both key authors and key users. Lucid 3 will meet (and to some extent guide [7]) specifications for the Standard for Descriptive Data (SDD) under development by the Taxonomic Databases Working Group (TDWG) of the International Union for Biological Sciences [7]. The SDD is being designed to accommodate the management of data while it is being generated with seamless conversion of data between Lucid 3 and other software for other purposes [3,7].

For the SRMIF numerous conceptual challenges and considerations must be overcome. One includes the terminology used to represent plant traits (=characters) and their variants (=character states).

Another challenge centers on issues related to key construction (top down versus bottom up). To address these issues we are constructing the key in 4 Phases, corresponding to 3 levels of the taxonomic hierarchy. Each phase will be released online in succession, which effectively allows “peer review” to occur. Phase 1 [1] is a key to plant families of the region. Phase 2 is a key to genera. Rapid progress in building the key’s underlying data matrix is made possible by using Lucid Score Analyzer [2] to assess the sufficiency of the data matrix before adding additional characters. Phase 3, a key to the species, will follow Phase 2. Lucid taxonomic data matrices from three offsite collaborators will be merged electronically.

Finally, Phase 4 will be add-on information, including images of species and selected character states, and notes on economic uses, conservation status, relevant literature, etc. Once finished, the interactive key will be linked to a federation of specimen databases from herbaria at the Universities of Northern Colorado, Colorado-Boulder, Colorado State, and Wyoming. The database management system for the online federation of databases is being developed by Aaron Reeves and collaborators using the DiGIR protocol for sharing data (http://digir.sourceforge.net), SuSE Linux and MySQL. The federation will contain georeferenced plant specimen data from the above herbaria, as well as specimens from our region at several other herbaria.

After completion of the SRMIF, annual (or even weekly) updates will be easy to produce.

3. Results and Discussion

Much groundwork for the 4 phases has been laid, including a downloadable synonymized checklist for the region [6], and progress on the interacting region-wide federation of specimen label databases (http://www.southernrockiesflora.org). Phase 1 [1] is an operational key to families. Phase 2, a mixed-rank key to families, most genera, and some species, is to be posted on the web in September 2004. A prototype Phase 4 treatment of the Heath family (Ericaceae) will be included in the September 2004 release. Phase 3, a key to all species, is scheduled for completion in 2006, but with a sizable portion of this to be released in 2005. Phase 4, is targeted for completion in 2006 or 2007.

Acknowledgements

Support comes from NSF grant DBI-0237149 to M. Simmons, R. P. Guralnick, R. Hartman, N. Snow, and T. Ranker. Additional support comes from the Colorado Native Plant Society and the University of Northern Colorado. Lucid help has been given by D. Sharp and the Lucid Development Team. A more complete listing of acknowledgements can be found in the “SRMIF checklist” [6].

References


CoCoLysis: A Web-Accessible Coiled Coil Protein Database with Analysis Tools

David Brinkmann,† Sai Nandoor,‡ Jugal Kalita,‡ Brian Tripet§ AND Robert Hodges†
sainandoor@hotmail.com, david.brinkmann@hp.com, kalita@pikespeak.ucsc.edu
brian.tripet@uchsc.edu AND robert.hodges@uchsc.edu

Abstract

Coiled Coils are simple quaternary protein structures that are used frequently in studies involving protein folding and design. They are also one of the most common oligomerization domains. It is estimated that 5-10% of the sequences revealed by genome projects encode coiled coil domains.

There are currently several online databases of protein sequences; however, no database exists online specifically for proteins with coiled coil domains. In this paper, we discuss a database designed specifically for proteins with coiled coil domains, along with the necessary application programs for stability and frequency analyses.

Extracting proteins with purported coiled coil domains manually from the ExPASy or any other protein database is a tedious process. Also, new entries are added to the databases on a regular basis. We programatically extract the protein sequences that contain coiled coils. The coiled coil parts of the protein sequences are also extracted. This data is loaded into a database that can be queried online. The information in the database can be updated to the latest additions by running an update script.

The source of data for the database is Swiss-Prot and EMBL. One source of data involves coiled coil domains that were previously identified. The other source of data is the full protein sequence database.

Figure 1: Heptad Positions in a Coiled Coil from which we identified the coiled coil domains.

1 What are Coiled Coils

A coiled coil is a oligomerization domain that is formed when two or more α-helices wrap around each other in a left-handed super coil. Coiled coils are found throughout nature, occurring in a wide variety of proteins and play an important role in biology. Two examples are kinesin [5] and myosin [6]. Both proteins are mechaenochemical proteins capable of utilizing chemical energy from ATP hydrolysis to generate mechanical force. The ability to perform this is due in part to the coiled coil [1]. Coiled coils are found to have hydrophobic amino acids spaced at every third and then every fourth residue within its sequence. A grouping of seven residues form a heptad repeat, designated abedfg where a and d positions are occupied by hydrophobic amino acids. This pattern can repeat, for example, 3 to 40 times depending on the particular coiled coil. In a coiled coil, the two α-helices bury their hydrophobic residues in the center of the coiled coil resulting in a super coil. These are depicted as positions a, a', d and d' in Figure 1. The super coil nature of the coiled coil also gives rise to interactions within the individual α-helices and between the α-helices in the super coil.
2 Why Coiled Coils

In order to study and predict protein structures in general, a first step may be to thoroughly understand one of the simplest protein structures: the coiled coil domain has only one type of secondary structure, the $\alpha$-helix, making interpretations of results less complex. Once we have acquired significant knowledge about coiled coils, the knowledge can be used to predict the relative stability and other features of previously unseen coiled coils and possibly more complex proteins. In particular, there are a number of advantages to studying the coiled coil domain:

- Abundant motif found in a variety of proteins
- There is only one type of secondary structure present, i.e., the $\alpha$-helix
- Only two interacting $\alpha$-helices are required to introduce tertiary and quaternary structure
- Diversity in length makes it an ideal system to test predictions
- All the non-covalent interactions that stabilize the three-dimensional structure of proteins are found in coiled coils
- It is experimentally easy to analyze structure and stability

Prior research has demonstrated that there are a number of possible factors that determine the stability of a coiled coil [2, 4, 7, 8]. Using these stability factors, a protein can be evaluated to find possible coiled coil domains within it. Once these domains are found, they can be further studied. Information about the domain’s composition and other statistics can be gathered and used to predict their presence in newly sequenced proteins.

3 CoCoLysis: Coiled Coil Database

Since we focus on the study of coiled coils, it is imperative to gather as many annotated sequences of coiled coils as possible for a thorough computational undertaking. A significant number of coiled coil sequences can be stored in a database allowing for probing queries to facilitate research in protein stability and structure prediction.

We have designed a PostgreSQL database\(^1\) to store a large number of coiled coil sub-sequences of proteins. The requirements placed on the database are simple, to start with. We want to allow searches in various ways; such as the name of proteins, species the protein is found in, the name of the gene that encoded the protein, and synonyms for the protein name. We also want to allow searches by amino acid sequence. Given a sequence of amino acids, we want to provide details about the closest matching coiled coil subsequences found in various proteins if any. The database keeps the structural and non-structural parts separately, but allows searches that correlate the two. The database also allows comparison of up to three search results. It also allows to search for clusters. A cluster is defined as starting and ending with three or more consecutive hydrophobic amino acids (Leu, Ile, Val, Met, Phe and Tyr) occupying the hydrophobic core positions with no more than one of these positions being occupied by anything else.

4 Source of Data

Designing a database involves deciding on the entities, the attributes of the entities, the relationships and the integrity constraints associated with the entities and relationships. We use PostgreSQL, an object-relational database management system. We use it in a pure relational manner. Hence, the next step is to convert the entity-relationship diagrams into relations or tables.

We use two sources of data which are related.

- **Swiss-Prot**: We decided to download all identified coiled coil subsequences from the Swiss-Prot database via SWall on the European Bioinformatics Institute (EBI) servers\(^2\). Here, we accept the coiled coil identification made by the Swiss-Prot Database.
- **STPR**: The second source of data is the entire protein database found in Swiss-Prot and TrEMBL\(^3\) database via the ExPaSY server\(^4\). Since STPR data is a collection of all proteins, a method for determining where in the proteins coiled coil regions appear is necessary. We employ our own algorithm for coiled coil identification. This algorithm is discussed later in the paper.

\(^1\)http://www.postgresql.com

\(^2\)http://www.ebi.ac.uk/Information/index.html

\(^3\)http://us.expasy.org/sprot/

\(^4\)http://us.expasy.org/
Both sets of data are pre-processed using our implementation of a program that considers factors that lead to stability of coiled coils to identify stable coiled coil regions. The data set derived from annotated coiled coils is referred to as the **Swiss-Prot** database and the data set derived from the entire Swiss-Prot TrEMBL database is referred to as the **STPR** dataset.

5 Coiled Coil Retrieval

To obtain the **Swiss-Prot** data set as mentioned above, we query the Swiss-Prot TrEMBL database for the coiled coil domain. It is a three-step process. First, a list of HTML links is extracted from the 2600 protein entry results to the query. Second, the HTML links are used to retrieve the Swiss-Prot entry page for each of the proteins. The contents of all the Swiss-Prot pages are parsed to find all the links to the COILED COIL (POTENTIAL) link. The final step uses the coiled coil linked pages to retrieve the page that has only the details of the coiled coil region. This process took over one hour and forty minutes using high-speed Internet connection. During this process, there were a number of “server time-out” errors. There were multiple attempts to get an error free run. This was not possible. The results from multiple entries had to be edited by hand in order to obtain a complete listing of the 2600 proteins.

6 Stablecoil Processing

Before any analysis can begin, the specific coiled coil regions of each sequence need to be determined. Coiled coils can be composed of multiple α-helices that wrap around each other. The individual coils are not necessarily aligned on the same heptad registry. To identify the coils on different heptad alignments, we process an implementation of the Stablecoil algorithm. Even though the Swiss-Prot data set has already identified purported coiled coil regions, the individual coils have not been identified. Using the Stablecoil algorithm, both Swiss-Prot and STPR data sets can be processed to determine specific coiled coil regions and the heptad registry shift in which they exist. The Stablecoil algorithm predicts the location and stability using alpha-helical propensity and stability coefficients as reported by [7, 8, 9]. By summing the residue scores over variable window widths and comparing the total score assigned to each amino acid to a known globular and cytoskeletal coiled coil containing sequences, the algorithm displays the region and probability (in kcal/mol) that a particular sequence will adopt a coiled coil conformation. Our implementation of the algorithm uses a 42 amino acid window width with a probability that the sequence is a coiled coil region set to 38 kcal/mol. Each sequence is processed seven times; once for each reading frame. Each amino acid has the combined helical propensity and stability coefficient applied to it based on its heptad registry position. The value of the amino acid position is determined by which heptad position it occupies in the heptad alignment. The amino acid position is set to one of three different values whether the amino acid is in the a, d, or one of the other five positions. After applying the values to the sequence, windowing is applied to locate coiled coil regions. This windowing process is repeated until all necessary updating of values have been performed. After the windowing process is complete, the regions that have at least 3 heptads with a value of greater than 38 are deemed to be coiled coil regions. These regions are then extracted from the sequences and saved along with the heptad registry positions with which it was found. When preprocessing is complete, all coiled coil regions in all the protein sequences are identified and each coiled coil sequence has a starting heptad offset assigned to it. These new sequences are placed in one of two new datasets that are used in this analysis. The first dataset, containing 2817 coiled coil sequences, is the Swiss-Prot data, having originally come from the Swiss-Prot coiled coil annotated database; the second dataset, containing 67358 coiled coil sequences, is the SPTR data, having been derived from the entire STPR database.

7 Navigational Aid

The database of coiled coils have a large number of entries. Hence, it is necessary to provide navigational aid to the users of the database. The database can be searched by criteria such as ExPASy name of the protein, ExPASy accession number, synonyms for the protein, gene name, species the protein is found, and a combination of these characteristics. However, a large database becomes difficult to search since the researcher using the system may not know what is
actually in the database, and whether it is biased in some ways. Hence, it is important to provide navigational aids to the user apart from a simple search form. Thus, we provide a way to navigate the database contents in a hierarchical manner. The user is able to navigate the database by species name or by protein name. For example, if someone is navigating by species, the user sees a list of species, in alphabetical order, that are in the database, and he can see how many proteins and how many coiled coil domain sequences are there for each of the species in our database. If the person wants to navigate by protein, the person sees the name of the proteins in an alphabetical order and then the number of entries in the database against each protein name in parenthesis. This helps the user get a feel for the contents of the database right away.

The system also allows one to compare the results of up to three searches. The database is queried for proteins matching each of several criteria as specified above. The user can save the results of up to two prior searches and compare results in various ways.

8 Stability Analysis

Once we have a protein or a coiled coil domain that comes out as a result of a search, we may be interested in determining how stable it is. Being able to determine protein stability is important because a minimum threshold of stability is required to initiate protein folding and stability is intimately involved in conformational changes and function of proteins [2, 3, 1]. To expedite this work, an analysis tool is needed to calculate the stability of an amino acid sequence. Stable Input, which is a part of our repertoire of analysis tools available with CoCoLysis was developed to help determine coiled coil stability over an entire sequence.

An HTML graphical user interface is available to provide input to Stable Input. The tool allows the biologist or biochemist the opportunity to choose a sequence from search results, set parameters, and perform calculations based on custom or default parameter values. The results are provided in the form of up to eight different graphs and a tab-delimited text file of sequence values in kilo-calories per mole (kcal/mol). The input from the HTML program is parsed and a CGI program written in Perl calculates the results. The calculations are based on either user inputs or program defaults. The user settable inputs are summarized below:

- Heptad Registry offset
- Window width
- Helical Propensity
- Hydrophobic core stability between a and d positions
- Intra-chain electrostatic interactions
- Inter-chain electrostatic interactions
- Hydrophobic Clusters

The details of the algorithms and formulas behind Stablecoil are not discussed in this paper due to lack of space. The Stablecoil tool is able to do the following analyses and present graphs:

- Total stability of the protein
- Hydrophobic stability in a and d positions
- Helical propensity
- Electrostatic interaction in e and g positions
- Chain length

Here, we show two graphs that the Stablecoil tool produces for a specific sequence of amino acids. Figure 2 shows the helical propensity of the windowed values for the individual amino acids. The graph shows regions where coil formation is favored. Since a single amino acid cannot form a protein, the windowing of the helical propensity shows the propensity for a region. The strongest region shown here is in the region approximately between 60 and 120, but to a lesser degree the entire protein shows a propensity to form helical structures. Figure 3 windowed or not, is one of the sparsest graphs generated. The E/G interactions are based on finding two oppositely charged (Lys, Glu, Asp, Arg, or His) amino acids in the e or g heptad positions.
9 Statistical Analysis

The database has associated tools to perform detailed statistical analysis of the two whole data sets: Swiss-Prot and STPR. It can also perform statistical analysis on the results of any search or any entry in the navigational hierarchy. We provide the appropriate Web interface to ask statistical queries. Some of these queries result in graphs that are produced by the GnuPlot software. Analysis produced include the following:

- Normalized length frequency over all coiled coils. Most are in the 42-50 range
- Number of Amino Acids in the a and d positions considering 6 and 7 heptads in the SPTR database
- Number of Amino Acids in the a and d positions considering 6 and 7 heptads in the Swiss-Prot database
- Normalized Amino Acid Distribution in Swiss-Prot
- Normalized Amino Acid Distribution in SPTR

For example, the Swiss-Prot and STPR data sets have a great variety of sequence lengths. A graph depicting this variety in total sequence length as obtained by our analysis tool is shown in Figure 4. We can get a similar graph at any level of hierarchy during the navigation of the contents of the database.

It is also possible to obtain frequencies of amino acids at any level during the navigation of the database or for the result of a search. For example, for the complete SPTR database, Figure 5 shows the relative frequency the amino acids appear in the a and d positions for both sets of data. The SPTR data shows that the a position is dominated by Leu, Ile, Val, Phe, Ala, and Tyr and in the d position Leu, Ile, Val, Ala, Phe, Met, and Tyr. The SPTR dataset shows that Ala competes with the hydrophobic amino acids in occurrence frequency. Other studies [7, 8] have found that Leu is most likely to be found in the a and d positions followed by the other hydrophobic amino acids with a strong showing of Ala in both the a and d positions. The strongest disagreement was in the frequency in which Met occurred. This study showed it was consistently one of the least likely hydrophobic amino acid to occur in the a and d positions, but in the other studies, Met was the third most likely hydrophobic amino acid to appear in the a and d positions. The stabilizing effect that clusters have on coiled coils has been proven experimentally.

Do long coiled coils have more clusters? If they, do how are they characterized? To answer this question the clusters found in all the sequences in both data sets can be examined by our analysis tool. A minimum sequence length of 42 amino acids or 6 heptads is examined and compared. The distribution

---

6 http://www.gnuplot.info/
of the normalized cluster length across all sequence lengths is shown in Figure 6. This figure shows the total number of clusters of length three or greater that appear in the various length sequences. The Swiss-Prot dataset has 5526 clusters and the SPTR dataset had 102718. We can create similar graphs for any search result or for any entry during navigation of the database entries. We observe that the Swiss-Prot dataset has a slight propensity for having fewer clusters in shorter sequences than that of the SPTR dataset. As the sequences get longer the cluster count for both sets of data falls off, but the fSPTR data diminishes more rapidly than that of the fSwiss-Prot data. While the SPTR dataset approaches no clusters counted beyond 12 heptads in length there is a relative consistency from length 12 through 19 heptads. Since the Swiss-Prot data comes from the coiled coil data set this data seems to suggest that clusters are important in longer coiled coils.

10 Conclusion

We have created a database of coiled coil domains of all proteins available at this time. We have two ways to obtain the coiled coil data. Our program also constantly monitors the Swiss-Prot and EMBL sites and any time any additions or updates are made it is reflected in our coiled coil database, keeping it up to date. The system displays the summary information that is current in the form of graphs and tables. The database has been designed for flexibility and future expansion.

References


A Mathematical Model for the Human Arterial System

Robert Carlson and Radu C. Cascaval
Department of Mathematics
University of Colorado at Colorado Springs
Colorado Springs, CO 80933

Abstract

We present a mathematical model for the pulse wave propagation in the human arterial system. The main feature we want to model is the one-dimensional nature of the wave dynamics in the tree structure, in the presence of variable characteristics along the vessels, such as non-uniform cross-sectional area and variable elasticity of the vessel walls. We describe the appropriate branching and boundary conditions which guarantee a well defined dynamical system. Moreover, numerical experiments illustrate the dynamics of individual pulses. Possible applications include the prediction of dynamic changes in the presence of stents or plaques in arteries, as well as changes in pressure due to modifications in stiffness of the arterial walls.

1. Introduction

The dynamics of the human circulatory system has been the subject of extensive study, both in the engineering and biomedical communities. Various mathematical models have been developed, to describe both local (individual vessels, [1], [5], [2]) and global dynamics (systemic tree, [3], [4]). The key feature of the arterial system is the coupling between the fluid and the elastic structure of the compliant vessels. Commonly, in the large arteries, the blood is treated as a Newtonian fluid (linear relation between shear stress and strain), hence the mass and momentum conservation are governed by the Navier-Stokes equations. The vessel are modeled as fluid filled tubes with linear elastic walls, while the fluid structure interaction follows a simple Newton’s law of motion. In this work we will use a similar description for the blood and for the arteries, allowing variable characteristics along the length of the tree, such as various radius and elasticity. Our focus is to establish a one dimensional simplified model that captures the main features of the pulse wave propagation along the systemic tree. While such a model cannot capture the complex dynamics due to the geometry of bifurcations or the presence of asymmetric occlusions (2D or 3D models would be a better fit), our primary goal is to create a model for the entire systemic tree, which can run in real time.

2. Mathematical Description

We start with the Navier-Stokes equations for describing the fluid motion in an axisymmetric domain. Let \((u,v)\) denote the longitudinal and radial components of the fluid velocity (in the \(x\) and \(r\) direction, respectively). Then

\[ u_t + uu_x + vv_r + \frac{1}{\rho} p_x = 0, \quad v_t + uv_x + vv_r + \frac{1}{\rho} p_r = 0 \]

\[ u_x + v_r + \frac{1}{r} v = 0 \]

Denote by \(\eta = \eta(x,t)\) the wall displacement at location \(x\) and time \(t\). Then the equation of the wall motion is

\[ \rho^w h \eta_{tt} = p^w - \frac{E h}{r_0(x)^2} \eta \]

where \(p^w\) is the transmural pressure at the wall, and the last term represents the circumferential stress.

2.1 1D Model

We exploit the fact that, in the large arteries, the lengths of propagation are relatively large compared with the radius of the vessel (that is \(\epsilon = R/L < < 1\)) and derive the long-wave approximation for the fluid-structure interaction system above [1]:

\[ \eta_t + \gamma \eta_x + \sigma \eta \eta_x - \alpha^2 \eta_{xx} = o(\epsilon^2) \]

The coefficients appearing in the PDE above are functions of the radius \(r = r(x)\) and elastic Young modulus \(E = E(x)\), which may be variable along the tree. The transmural pressure \(p = p(x,t)\) is proportional to the wall displacement, more precisely

\[ p(x,t) = \frac{2}{r(x)^2} E(x) \eta(x,t), \]

hence it satisfies a similar equation as \(\eta\).

The inflow boundary conditions (at the ascending aorta) are taken to be of the form \(p(0,t) = t^2 \exp(-\delta t) t \geq 0\), where \(\delta\) is related to the duration of systolic phase, and
as out ow conditions we choose Neumann boundaries: \( p_x(v, t) = 0 \) for all terminal vertices \( v \). The branching conditions are chosen on the premises that pressure is continuous across branching and the incoming \( u_x \) equals the outgoing \( u_x \). This translates into

\[
p(v^-) = p(v^+), \quad p_x(v^-) = \sum_{u^+} p_x(v^+)
\]

which is the analog of Kirchhoff’s law for electrical circuits.

### 2.2 Numerical Simulation

We consider a tree structure (see below) which accounts for the major arteries in the human circulatory system (46 large vessels), see [4]. Each vessel has an initial radius and a terminal radius, which indicate tapering effects are present in the model.

Then the numerical computation reduces to a system of 46 equations on a finite interval (say [0,1]), with initial and boundary conditions. The model equation for the pressure \( p \) on each (scaled) edge can be rewritten in the form

\[
(I - \alpha^2 \partial_x^2) p_t = -\gamma p_x - \sigma p p_x, \quad p(0) = f, \quad p(1) = g
\]

or, using the resolvent kernel for the \(-\partial^2\) operator,

\[
p_t(x,t) = \int_0^1 \partial_x G(x,y) [\gamma p(y, t) + \sigma p(y, t^2)] dy + c(f, g)
\]

where the explicit term \( c(f, g) \) is obtained from an integration by parts. We choose discrete sampling points on each edge \( i, x^1_i, \ldots, x^N_i \), and denote \( p(x^i_k) = p^{(i)}_k \). The branching conditions above become (in the discrete version), \( p(v) = \frac{1}{3} (p_{N-1} + p_1^2 + p_3^3) \), where the vertex \( v \) has parent edge labelled 1, and daughter edges labelled 2 and 3.

The numerical algorithm uses a fourth order Runge-Kutta scheme for the time discretization and a gaussian quadrature for the integral term. The boundary conditions used are derived from Kirchhoff’s law with the precomputed values of pressure on adjacent edges; the exception is the initial edge (where the pressure value is determined by the in ow condition) and the terminal edges, where discretized Neumann conditions apply.

### 3. Summary and Conclusions

We have established a 1D tree model to describe the pulse wave propagation in the arterial tree. Numerical simulations have been designed to take into account in ow conditions (at the beginning of ascending aorta) which follow physiological characteristics (and can be extracted from real MRI data). The out ow boundary conditions are difficult to match the real ones, since it is hard to collect data at each site (organ). The variable radius and elasticity can be used to study changes in pressure due to stents, partial occlusions, or changes in stiffness of the vessel walls. Our model does not capture possible reflected waves, which in the real system account for a large portion of the dynamics at the smaller arteries level. Future work will consist of introducing the reflection of waves in our model as well as the wave propagation in the terminal trees, and compare our numerical results with physiological data.

### References


Character Based Phylogeny Inference in the Grid Computing Environment

Zhijie Guan
Department of Computer and Information Sciences
University of Alabama at Birmingham
zhijie@hpcl.cis.uab.edu

Recently the availability of efficient data collection tools in information technologies flood biological researchers with a huge amount of genetic information. Phylogenetic inference could have made greater progress if there were effective solutions to evolutionary tree search and evaluation. This paper demonstrates how to bring into reality a high-performance parallel character based approach for phylogeny inference within the Grid computing environment. Accordingly, resource scheduling, load balancing, and asynchronous communication are adopted to achieve efficient parallel branch-and-bound search.

1. Introduction

In recent years, advances in information technologies have enabled biologists to explore issues in life sciences at scales greater/finer than ever before. The availability of efficient data collection tools present researchers with a huge amount of genetic information from which phylogenetic inference can be performed. The evolutionary relationships among taxa are usually described as labeled bifurcated tree. Though a set of algorithms have been developed to address phylogeny inference based on different evaluation metrics [6], the phylogeny reconstruction problem is still computationally complex (NP-hard) because the number of potential solution trees increases exponentially with the number of taxa included in the phylogenetic analysis. Facing the computational difficulty of phylogeny reconstruction problem, scientists develop two kinds of approaches with two different goals. One approach is to use fast but inaccurate search algorithms to find near optimal solutions in limited time. The other resorts to parallel methods to reduce the total analytical time. In this paper, we present a high-throughput architecture of the parallel branch-and-bound [11] approach for phylogeny inference, which takes advantages of the potentially unlimited computing resources in the Grid environment.

2. Related Work

Phylogeny inference is the process to find the particular phylogenetic trees that best represent the evolutionary history of the analyzed taxa, according to a specified optimality criterion. Using different optimality criteria, inference methods can be categorized into distance-based methods (like neighbor-joining method [4] and UPGMA [7] method) and character-based methods [5]. Numerous character-based methods have been proposed with various search schemas, like exhaustive search, branch-and-bound search, and stochastic search [6]. Serial are most of the phylogenetic software packages, like PAUP* [10], PHYLIP [1], and fastDNAML [3], employing branch-and-bound search for small datasets analysis, and stochastic search for large datasets inference. Although a few approaches [2,8] touch upon parallel programming and execution of phylogenetic algorithms, many of them have been narrowed to tight-coupled parallel applications. Stamatakis etc. [9] proposed a seti@home-like RAxML approach that is promising on the loose-coupled parallel inference but only with heuristic search methods.

3. System Architecture and Implementation

3.1 Parallel Branch-and-bound Search

![Figure 1 System Architecture](image)

Branch-and-bound search [11] could be applied to phylogenetic analyses with any criteria that is non-decreasing as taxa added in a stepwise manner. It constructs the candidate tree by adding taxa to an initial tree gradually till all taxa are involved. After each step of the addition, the generated tree is evaluated against the optimality criteria. If the value of the optimality criteria exceeds a pre-set bound, the tree and all of its derived trees (trees obtained by adding branches to this tree) are pruned without further consideration. The pre-set bound is updated whenever a better value of optimality criteria is found during the evaluation of a complete tree. Parallel branch-and-bound search distributes the tree generation and evaluation job to multiple processes. A special
communication schema is built up to update the bound among processes and to collect the best solutions for the inference.

This parallel branch-and-bound search approach is briefly described as in Figure 1. The inference job is distributed across the network/Grid area and performed on multiple computers simultaneously. A master node is employed to partition the inference job into small sub-jobs, put sub-jobs into Condor [12] job queue, update the bound of evaluation metric for every slave nodes, and to record the best tree ever found during the search. Condor pool manager [12] is a server that dynamically catches available computing resources in the network/Grid and dispatches queued jobs onto them. Upon receiving a job from the Condor pool manager, a slave node begins by registering itself with the master node, fetching initial tree and criteria bound from the master node, and performing the branch-and-bound search in local environment. Whenever a score better than the bound is identified during the inference, it will be reported to the master node that in turn broadcasts the new bound to all slave nodes in the pool.

### 3.2 Implementation

In order to achieve an efficient and flexible implementation, two key challenges have been identified, which are attempted to be addressed in this research project. First, the problem of resource scheduling and load balancing among slave nodes are solved by applying a First-In-First-Out (FIFO) scheduling schema on the Condor task queue. Condor can automatically locate idle computers on the network and assign jobs to them. What’s more, Condor can find resources not only on isolated networks, but also in the potential global Grid computing environment, where this approach can explore more computing power as well as demonstrate better scalability. Secondly, the efficiency of the pruning algorithms is dramatically increased by using an asynchronous communication pattern among master-slave nodes. The communication between master and slave nodes is in the charge of an individual thread in each slave node, which updates the bound in the pruning thread in real-time. This reduces the communication overhead in the thread for tree generation.

### 4. Results

Tests on a synthesis dataset and a real dataset were carried out to evaluate the performance of this approach. A Condor pool with 32 nodes was used. The synthesis dataset, which needed more than 10 hours to perform the branch-and-bound search in sequential mode, needed less than 2 hours to finish, with a peak usage of 8 nodes simultaneously. The inference time of the real dataset (40 taxa with 9476 base pairs each) was also reduced from 181 hours in sequential mode to 22.5 hours with a peak usage of 32 nodes.

### 5. Conclusion and Future Work

A high-throughput architecture of parallel branch-and-bound search approach for phylogeny inference in the Grid computing environment is presented in this paper. Condor job queue and asynchronous communication mechanism are applied to achieve better performance and flexibility. Initial experiments with two datasets demonstrate the promising ability of this approach. Future work includes, but is not limited to, parallelizing more heuristic search algorithms in the Grid environment and increasing the accuracy of results.

### 6. References


A Web-Based Tool for Modelling and Simulation of Ion Channels

Mona Soliman Habib AND Jugal Kalita
Department of Computer Science, University of Colorado, 1420 Austin Bluffs Parkway
Colorado Springs, Colorado 80917 USA
msoliman@cs.uccs.edu, kalita@pikespeak.uccs.edu

Abstract

Background: Ion channels are proteins in the cell membrane that facilitate the diffusion of selected ions through biological membranes. Measuring ionic current has been made possible using the giga-seal patch-clamp technique [13]. Modelling and simulation provide a better understanding that complements experimental results. Several tools are available. New modelling standards based on markup languages are being developed in order to promote collaborative research and model sharing. This paper presents a web-based channel simulation tool that uses a database of models described by the NeuroML language [6]. The tool currently focuses on the simulation of voltage-gated ion channels and can be used as a prototype for a more generalized, standard-based and Web accessible neuronal modelling and simulation engine.

Results: The Ion Channel Simulation tool reads NeuroML models from either a database or a user-supplied file and simulates the forward and backward rates variations with voltage, the channel open probability and time constant variations with voltage, and the gate and channel conductance and channel current density variations with membrane voltage and time. Tests performed include the Na and K squid axon model by Hodgkin and Huxley [8], and the Ca3 hippocampal pyramidal neuron model by Traub et al. [14].

Conclusion: The Ion Channel Simulation tool is a prototype for a user-friendly neural simulation engine that promotes collaborative research and model sharing via a Web-based interface. Future research directions include exploring other modelling techniques and studying the properties of different ion channels as a basis for understanding the kinetics and electrochemical properties of neurotransmitters. The ultimate research goal would be to explore the use of modelling techniques in order to develop a reduced and reliable data representation for brain electrochemical activity.

1 Background

Ion channels regulate electrical activities in the nervous system and control muscle movements. Therefore, understanding the electrical and stochastic gating properties of ion channels is a fundamental question in neuroscience. The giga-seal patch-clamp technique [13] allows measurement of ionic currents flowing through single channels in the cell membrane. Modelling ion channels serves several purposes such as capturing one or more features of the natural systems in a simplified context, predicting the structure and permeation mode of ion channels, and helping in the design of artificial channels and probing sensors [15]. Kuyucak et al. [10] and Chung [4] discuss various approaches to modelling of ion transport in membrane channels, including continuum theories, Brownian dynamics, and classical and ab initio molecular dynamics. Hodgkin and Huxley [8] present a mathematical model for voltage-gated ion channels. Other techniques for modelling the stochastic gating of ion channels, such as using a two-state Markov process, Monte Carlo simulation, and methods for simulating multiple independent ion channels are described in [8]. With increased understanding of the structure and function of ion channels, reliable and accurate modelling may help in the specification of a reduced set of data points that describes the channels conductance accurately and efficiently. A reduced data set would pave the way to the development of simplified and efficient computational algorithms for neuroscience data analysis and knowledge discovery.

A number of software packages for neuronal modelling are available, such as Genesis [2], Neuron [7], Neurosim [5], and Catacomb2 [3]. While being powerful tools for modelling and simulation of neuronal processes and networks at various levels, these tools require porting to the user machine and a steep learning curve. Moreover, each of these tools uses its own scripting language and parameter definition file formats, thereby limiting the possibility of sharing mod-
els among different tools.
As the needs for sharing research results and developed models among scientists increases, several efforts for creating standard modelling languages take place. CellML [11], NeuroML [6], and SBML [9] are markup languages used to define biological and neurological systems in a simulator-independent representation scheme and aim to provide a basis for collaborative modelling efforts. Efforts to support one or more of these markup languages on existing tools are underway at different stages of development. The use of databases to store developed models is another step towards collaborative research. The Modelers Workspace Channel Database (ChannelDB) [1] and NeuronDB [12] are examples of database models that can be linked to Neuron or Genesis simulation tools.

In this paper, we present a prototype for a web-based ion channel simulation tool based on NeuroML. The current implementation supports the simulation of voltage-gated channels based on the Hodgkin-Huxley mathematical model [8]. The tool may be extended in order to support other channel types and modelling techniques, and to provide the ability to generate NeuroML models based on user-defined parameters. As a web-based tool does not require porting of software packages to the user machine, this prototype may serve as a basis for a more generalized neuronal modelling and simulation tool that provides an easy mechanism for sharing models described in one or more of the emerging standard markup languages.

2 Materials and Methods

The Ion Channel Simulation tool is a prototype for a Web-based neural simulation engine that is based on some of the emerging standards in computational neuroscience. The program reads models from either a database or XML files, where models are defined using the NeuroML markup language. The server backend is an object-oriented software package written in Perl. The prototype implements a subset of the ChannelPkg subschema of NeuroML, focusing on the classes that define voltage-gated ion channels.

2.1 Mathematical Model

The mathematical model is based on the Hodgkin-Huxley mathematical formalism [8] for voltage-gated ion channels. The model is based on the electrical equivalence between an ion channel and an RC electrical circuit with reverse potential.

In the model in Figure 1, \( C_m \) represents the cell membrane capacitance, \( g_m \) represents the gate’s conductance (\( g_m = \frac{1}{R_m} \)) where \( R_m \) is the gate’s resistance, \( E_i \) is the reverse potential, \( V \) is the membranes potential, and \( I_c \) and \( I_m \) are the current densities.

For a fixed membrane voltage \( V \), the channel current is given by:

\[
I_m = C_m \frac{dV}{dt} = -g_m (V - E_i)
\]

Let \( \alpha(V) \) be the rate at which the channel is open at a membrane voltage \( V \), and \( \alpha(V) \) be the rate at which the channel is closed at the same voltage \( V \). Let \( g \) be the gate conductance when the membrane voltage is \( V \). The generalized Hodgkin-Huxley model defines \( g \) as:

\[
g_m = g_{max} n^p
\]

where, \( p \) is a power specific to the gates type, and \( n \) is a probability given by:

\[
\frac{dn}{dt} = \alpha_n(V)(1 - n) - \beta_n(V)
\]

\[
\frac{dn}{dt} = \frac{1}{\tau_n(V)} (n(\infty) - n)
\]

where, \( n(\infty) \) is the probability of the channel (gate) being open at membrane voltage \( V \), and \( \alpha_n(\infty) \) is a time constant. \( n(\infty) \) and \( \tau_n(\infty) \) are expressed as follows:

\[
n(\infty) = \frac{\alpha_n(V)}{\alpha_n(V) + \beta_n(V)}
\]

\[
\tau_n(\infty) = \frac{\alpha_n(V)}{\alpha_n(V) + \beta_n(V)}
\]

Two solutions for the gates conductance differential equation are implemented. The first solution is an incremental numerical one where,

\[
n(t + \alpha_t) = n(t) + (n(\infty) - n(t))(1 - \exp(\frac{\alpha_n(t)}{\alpha_n(\infty)}))
\]

The second solution uses the exponential Euler formula leading to the solution:

\[
n(t) = n(\infty) - (n(\infty) - n(0))\exp(\frac{\alpha_n(t)}{\alpha_n(\infty)}
\]

where, \( n(0) \) is the probability of the channel being open at membrane voltage zero.

\[
n(0) = \frac{n(0)}{\alpha_n(0) + \beta_n(0)}
\]
For a channel with multiple gates, the overall channel conductance is given by:

\[ g_m = g_{max} \prod p_i \]

for all gates i. The gate current density (current per unit area) is given by:

\[ I_m = g_m (E_i - V) \]

The total current density of a multi-channel membrane model is the total of the individual channels current densities. Finally, given the channels current density, the membrane’s action potential is obtained by solving the differential equation:

\[ I_m = C_m \frac{dV}{dt} = -g_m (V - E_i) \]

### 2.2 Channel Database Schema

Ion channel models are stored in a Postgresql database, ChannelDB. This database consists of one main table, Channel, using a schema that is consistent with that used in the Modelers Workspace Channel Database project [1], an extension of the GEneral NEural SImulation System (Genesis) project [2]. Schema of the Channel table is as follows:

<table>
<thead>
<tr>
<th>id</th>
<th>varchar(64)</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>varchar(32)</td>
</tr>
<tr>
<td>ion</td>
<td>varchar(32)</td>
</tr>
<tr>
<td>keywords</td>
<td>varchar(255)</td>
</tr>
<tr>
<td>description</td>
<td>varchar(255)</td>
</tr>
<tr>
<td>xml_document</td>
<td>text</td>
</tr>
</tbody>
</table>

### 2.3 Model Class Definitions

The model class definitions of the Ion Channel Simulator tool are consistent with the NeuroML [6] schema in order to ensure compatibility and provide a basis for a neuroscientific simulator tool that adheres with the standardization effort in this field. The classes currently supported in the Web-based simulation tool are a subset of the NeuroML Channel subschema which describes voltage-gated ion channels. Support for other channel types can be easily incorporated using the base structure of the simulator.

A description of the classes used in the ion channel simulation tool follows.

**ChannelModel:** The ChannelModel class describes the higher level model attributes as follows:

- author: The author name.
- description: A short description.
- date: The model development date.
- literatureReference: The reference where the model is described.
- eRest: The nominal resting potential of the cell for which this channel was developed.
- keywords: Space separated list of keywords.
- notes: Additional notes.
- uniqueID: A unique identifier used by the database to allow multiple channels to have the same short names.
- channels: A set of channel definitions.

**Channel:** The Channel class is the base class for channel definitions. It describes base attributes and methods. The class has the following attributes:

- name: Channels name.
- class: Subclass that best describes the channels type.

**HHChannel:** The HHChannel class describes a Hodgkin-Huxley channel with the following attributes:

- eRev: The nominal reverse potential across the channel.
- gmax: The maximum value for the conductance variable.
- permeantSpecie: The permeation ion.
- gates: A set of gate definitions.

**HHGate:** The HHGate class is the base class for gate definitions. It describes base attributes and methods. The class has the following attributes:
• name: Gates name.
• class: Subclass that best describes the gates type.
• power: Power used to compute the gates conductance.

HHVGate: The HHVGate class describes a voltage-gated channel using the following attributes:
  • forwardRate: Alpha rate variable in Hodgkin-Huxley equations.
  • backwardRate: Beta rate variable in Hodgkin-Huxley equations.
  • instantCalculation: Boolean. If true, perform an instantaneous calculation of the gate activation variable, setting it equal to alpha/(alpha+beta).
  • timeUnit: Time unit used with the rate variables.
  • useState: Boolean. If true, use forwardRate, backwardRate for state variables tau, minf instead of rate variables alpha, beta.
  • vmax: Maximum voltage expected to be applied to the gate.
  • vmin: Minimum voltage expected to be applied to the gate.
  • voltageUnit: Voltage unit used for the rate variables.

HHRate: The HHRate class is the base class for rate definitions (forward and backward). It describes base methods for rate calculations. The class has the following supporting attributes:
  • direction: Forward or Backward rate.
  • class: Subclass that best describes the rate calculations type.

EquationHHRate: The EquationHHRate class describes a rate as a mathematical expression given by the attribute:
  • rateExpression: Equation used to calculate the rate.

ParameterizedHHRate: The ParameterizedHHRate class describes a rate in terms of six generalized parameters. The parameters are used to compute any Hodgkin-Huxley rate using the following expression:

\[ \text{rate} = \frac{(A + BV)}{(C + D \exp\left(\frac{E+V}{F}\right))} \]

The ParameterizedHHRate class has the following attributes:
  • A: Variable A in the expression above.
  • B: Variable B in the expression above.
  • C: Variable C in the expression above.
  • D: Variable D in the expression above.
  • E: Variable E in the expression above.
  • F: Variable F in the expression above.
  • timeUnit: Time unit used with the rate variables.
  • voltageUnit: Voltage unit used with the rate variables.

TabulatedHHRate: The TabulatedHHRate class allows a gate’s forward or backward rate to be specified using a table of rate values at equally spaced voltage (or concentration) points. The class has the following attributes:
  • fillPoints: Number of points to use if the rate table is expanded with interpolation.
  • interpolation: Type of interpolation to use between table entries.
  • rates: Two-dimensional array containing rate values.
  • xmax: Voltage (or concentration) corresponding to the last point.
  • xmin: Voltage (or concentration) corresponding to first point.
  • xpts: Number of data points in the rates array.

3 Results

The following is a sample set of simulation results for the Hodgkin-Huxley Na+ squid axon channel model described in [8]. The simulation results obtained using this tool are consistent with those documented in literature. The sample simulation is generated over a 6ms time simulation period with a 6µs time step and a 3mV membrane voltage step. The variation of the channels action potential with time is not yet implemented.

The ChannelDB database includes models for other channels such as the Ca3 hippocampal pyramidal neuron described in [14]. The source of the NeuroML examples included in the database is the Modelers Workspace Channel Database [1] Website.

4 Conclusions

This paper presents a web-based tool for voltage-gated ion channel simulation based on the Hodgkin-Huxley mathematical model [3], the Modelers
Figure 2: Activation gate forward and backward rates as they vary with membrane voltage.

Figure 3: Activation gate open probability (minf) and time constant (tau) as they vary with membrane voltage.

Figure 4: Activation gate conductance as it varies with voltage and time.

Figure 5: Inactivation gate forward and backward rates as they vary with membrane voltage.

Figure 6: Inactivation gate open probability (minf) and time constant (tau) as they vary with voltage.

Figure 7: Inactivation gate conductance as it varies with voltage and time.

Figure 8: Na squid channel conductance as it varies with voltage and time.

Workspace Channel Database [16], and the NeuroML modelling markup language [2]. The tool constitutes a prototype for a user-friendly neural simulation engine that promotes collaborative research and model sharing via a Web-based interface. Using standard modelling languages such as NeuroML frees the user to focus on the models description in a simulator-independent environment. This project serves as a proof of concept towards the implementation of a generalized standard-based simulation tool.
For future research, we would like to explore other modelling techniques and study the properties of different ion channels as a basis for understanding the kinetics and electrochemical properties of neurotransmitters. The ultimate research goal would be to explore the use of modelling techniques in order to develop a reduced and reliable data representation for brain electrochemical activity. Such representation could possibly lead to the development of simplified and efficient computational algorithms for neuroscientific data analysis and knowledge discovery.

References

NOVEL SOYBEAN ANTIBIOTICS

Tamara J. McEwen, and Daniel M. Zurek, PhD.
Department of Biological Sciences, Pittsburg State University, Pittsburg, KS 66762
timcwen@yahoo.com, dzurek@pittstate.edu

ABSTRACT

The soybean xyloglucanase gene identified in GenBank as pBRU1 is known to encode a glucanase protein. Initial bacterial expression studies have provided evidence that the protein product of this gene exhibits antibiotic properties. pBRU1 DNA was amplified by PCR from soybean epicotyl cDNA using primers for the complete structural region of the gene and cloned into the pYES2.1 TOPO TA yeast expression shuttle vector (Invitrogen). One successful transformant from this shuttle vector was identified by restriction mapping and sequencing from a library of E. coli BL21 non-expressing transformants. This recombinant plasmid was then transformed into the S. cerevisiae INVSc1 strain (Invitrogen) in order to produce sufficient quantities of protein for purification and characterization purposes. Protein expression, induced by the addition of galactose and raffinose in the absence of glucose with this system, has allowed us to identify low levels of recombinant protein expression via purification using the 6X-His tag on a nickel column. Antimicrobial effects of the protein will be determined by culturing a variety of pathogenic organisms in liquid and solid media containing varying concentrations of the recombinant xyloglucanase and comparing antimicrobial activity against a bank of standard antibiotic compounds.
Applied Sampling Methods for Estimating the Total Number of SNPs in Human Chromosome 19

Leigh Quiogue¹, Christopher Chee²

¹Department of Mathematics & Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, USA
²Department of Neuroscience, UCLA, Los Angeles, USA

Introduction:
Single Nucleotide Polymorphisms (SNPs) are small variations that account for most of the genetic diversity among humans. SNPs occur, on average, about 1% of the time in the human population and are found in both coding and non-coding regions of the DNA sequence. Today, researchers investigating both Mendelian and complex diseases utilize SNPs because they are highly informative and numerous throughout the genome. SNPs are the most abundant class of DNA polymorphisms, and SNP arrays for genotyping require very small amounts of DNA. Furthermore, newly developed genotyping and sequencing technologies, such as PSQ 96MA™ and the Affymetrix dChip, make it faster and less expensive for scientists to analyze thousands of SNPs. The number of newly discovered SNPs increase daily as more and more individuals are genotyped. However, the total number of SNPs in the human genome is currently unknown and estimated totals for each chromosome vary among different databases, thus making it difficult and time-consuming for scientists to choose SNPs for their research studies.

Sampling Scheme:
For our study, we estimated the total number of SNPs on chromosome 19 in order to get a rough idea of how much information is available for access. We focused on chromosome 19, which has the highest gene density of all the human chromosomes. Chromosome 19 contains over 60 million base pairs and is known to house genes linked to atherosclerosis, liposarcoma, and deafness. Furthermore, due to chromosome 19’s small size (about 63.8 Mb), we were able to complete our estimation in less time and cover it’s entire length. To estimate the total number of SNPs for chromosome 19, we performed stratified random sampling and obtained a sample of 270 genes from the 1,320 known genes. We then enumerated the total number of SNPs in the 270 genes using the UCSC Genome Browser, July 2003 Build. We obtained the list of known genes from previous research done on chromosome 19 by Grimwood et al. We then defined the four strata according to gene size as follows and filtered the genes into each one:

- Stratum 1: 250K to 300K+ bp
- Stratum 2: 150K to 224,999 bp
- Stratum 3: 75K to 149,999 bp
- Stratum 4: 0 to 74,999 bp

There was 1 gene in stratum 1, 7 in stratum 2, 34 in stratum 3, and 1,278 in stratum 4. The stratification was done according to gene size to compensate for the fact that larger genes have more SNPs than smaller genes.

Using a simple counting program, we were able to obtain accurate SNP totals for each gene and stratum. The total number of SNPs in stratum 1 is 911 and the total for stratum 2 is 3,347,999 with mean=478.2857, σ=132.0842, and standard error=49.92314. For stratum 3, we took a simple random sample of 30 genes (N = 34) and obtained an estimated total of
7,759.9322 with mean=228.2333, σ=94.66554, and standard error=17.28348.
Finally, for stratum 4 we took a simple random sample of 232 genes (N = 1,278) and obtained an estimated total of 65,161.47546 with mean=50.98707, σ=67.30865, and standard error=4.419029. Using the estimated SNP totals from each stratum, we estimated the total number of SNPs in chromosome 19 to be 77,180.40756 with a standard error of 71.625649. The 95% confidence interval for our estimate is [77,040.02129, 77,320.79383].

There are other sampling methods (cluster, simple random, etc…) that can be used to estimate SNP totals as well. However, we chose stratified random sampling as the best design. With a D_eff value of 0.426, the data confirms that the stratified scheme gave a more precise estimation per observation unit than a simple random sample. One important note is that there were only 3 strata taken into consideration since stratum 1 consists of only 1 member. But even after it was included in stratum 2, the D_eff was slightly less, 0.399, which still confirms that a stratified random sample provided a more precise estimate.

Hypothesis Tests:
Using our estimated totals, we tested whether or not scientists’ claim of one SNP existing every 100-300 bp is a reliable estimate. We performed three t-tests (pertaining to stratum 2, 3, and 4) for a single mean when σ is unknown. The results were as follows:

Test 1: Stratum 2, N = 7, n = 7
X SNP = 478.2857, SD = 132.0842, s = 49.923
X size = 174,316.143 bp
H0: μ = 871.581
Hα: μ1 < 871.581
t = -7.878 with p < .001

Test 2: Stratum 3, N = 34, n = 30
X SNP = 228.233, SD = 94.666, s = 17.283
X size = 97,366.44 bp
H0: μ = 486.832
Hα: μ1 < 486.832
t = -14.963 with p < .001

Test 3: Stratum 4, N = 1,278, n = 232
X SNP = 50.987, SD = 67.308, s = 3.347
X size = 16,874.604 bp
H0: μ = 84.373
Hα: μ1 < 84.373
t = -9.975 with p < .001

For all three tests at the α = 0.05 level with n-1 d.f., we rejected the null. Based on our data from the UCSC database, we have evidence to suggest that the estimate of finding one SNP every 200 bp is not true for genes on chromosome 19 that fall in each of our strata. To test whether the UCSC database provides more or less SNPs than expected based on the claim, one should perform 1-sided t-tests. Here, we have simply shown that there appears to be a difference and that there is statistically significant evidence to reject the claim.

These analyses for chromosome 19 provide an approximation for the total amount of markers available for study. Most researchers want to know how many SNPs lie in their gene or region of interest. However, SNP enumeration is tedious due to overwhelming amounts of information scattered among various databases. We hope that in the future, SNP information for all of the human chromosomes will be available from one database—thus making it readily and easily accessible for use and study.

Contact e-mail: leighq@mednet.ucla.edu
Mailing address:
University of California, Los Angeles, Department of Human Genetics
Gonda Center, Room 5309
695 Charles E. Young Drive, South
Los Angeles, CA 90095
Analysis of Nasal Epithelium Gene Expression in Cystic Fibrosis

Thami Rachidi 1, Frank Accurso 2, Krzysztof Cios 1,3,4
University of Colorado at Denver and Health Sciences Center 1, Denver, CO
The Children’s Hospital 2, Denver, CO
University of Colorado at Boulder, Boulder 3, CO
4cData LLC 4, Golden, CO
trachidi@carbon.cudenver.edu, Accurso.Frank@tchden.org, Krys.Cios@cudenver.edu,

Our analysis is a part of an on going study of CF gene expression among patients at Children’s Hospital, Denver. Investigators wish to learn whether sampling nasal epithelium could function as a less-invasive surrogate tissue in biomarker development for lung pathology in CF. We use microarray technology to investigate the utility of nasal epithelium gene expression in CF diagnosis. One of the core applications of microarray technology is to detect sets of genes responsible for genetic diseases. Typical methods use t-test to compare experiments across conditions in order to find differentially expressed genes in gene expression data. The t-test approach calculates the p-value for each gene across all experiments, and a multiple testing method may be employed to reject the null hypothesis. The problem with the t-test approach is that it assumes that the distribution of the mean being tested is normal. This is not true for many gene expression data. To overcome this limitation of the t-test, this study uses Kolmogorov-Smirnov test (KS-test) [1], a non-parametric procedure, in which the data drives the distribution, to find the genes that vary across conditions. Another important step in microarray data analysis is normalization of the data in order to reduce noise in analysis and variability. Several normalization methods have been proposed, however, their impact on the power tail distribution has not been fully evaluated. We investigate the effect of four normalization algorithms on the genes that varied in nasal epithelium gene expression data. These normalization procedures are: Affymetrix Microarray Suite 5.0 (MAS 5.0) [2], Robust multichip analysis (RMA) [3], DNA-Chip (DChip)[4] based on the perfect match only, and DChip [4] based on perfect match-mismatch. The data used consists of the expression of 12,225 genes in CF and control samples. Each of these two groups contains 9 replicates. These data was acquired using Affymetrix GeneChips platform.

Our method of analysis consists of applying the four normalization algorithms to the data. Then, In order to increase the power of our statistical tests, two filtering methods are used to eliminate the irrelevant genes:
(i) Low variance filter [5]: the variance, $s^2$, for each gene across all experiments is calculated. The null hypothesis is that the gene hasn’t varied across the experiments. Then, the statistic $(N-1) \frac{s^2}{\text{median}(s^2)}$, where N is the number of experiments, is computed. The p-value for this statistic is calculated and adjusted using the False Discovery Rate method (FDR). FDR is set to 0.1. Genes having p-values less than the critical value are kept and the rest is eliminated.
(ii) Absent detection calls filter: genes that have absent detection calls in at least 17 experiments are eliminated from the data set.

The effect of normalization method on the distribution tail is studied by calculating KS-test for each condition, normal and CF, and p values are computed. FDR is used to adjust for multiple comparisons [6] by setting an acceptable proportion of false positives for the experiments. In this case, the p values are ranked in increasing order, then sequentially compared to an adjusted cutoff proportional to FDR. The null hypothesis is rejected if the observed p value is greater than the adjusted cutoff.

Results from our statistical tests indicate that the number of differentially expressed genes varies with the normalization algorithm. In addition, for both t-test and KS-test, FDR is high for all normalization methods, and the minimum FDR at which differentially expressed genes were detected is obtained when data is normalized with MAS 5.0.

We conclude that the choice of normalization greatly affects the differentially expressed genes, and MAS 5.0 is relatively better when compared to other algorithms. However, because of the high FDR, we suspect that there is no difference between the two groups CF and control.

References


The First Biotechnology and Bioinformatics Symposium:
A Community and Academic Forum

INDEX

Abel, Barrett .................. 11  Hu, Yanhui .......................... 1
Accuro, Frank .................. 92  Isakov, Steven ...................... 1
Anderson, Daron ............... 66  Iyriboz, Tunc ..................... 4
Baker, Kristyn ................. 68  Jaimpojamarn, Sittichai ...... 61
Bartlemez, Stephen ............ 56  Kalita, Jugal ..................... 30, 50, 68, 73, 83
Boult, Terrance ...............  4  Karim, M E ........................ 25
Brasher, Jeffrey Wayne ....... 71  Keselj, Vlado .................... 61
Brinkmann, David ............. 73  LaBaer, Joshua .................. 1
Brugge, Joan ................... 1  Lakhotia, Arun ................... 25
Carlson, Robert ...............  79  Longanantharaj, Raja ........ 25
Cascaval, Radu C .............. 79  Los, Layne ........................ 36
Cercone, Nick .................. 61  McEwen, Tamara J ............. 89
Chandrashekar, Kaushal ...... 30, 50  Meyer, Lane ................... 11
Chee, Christopher .......... 90  Nandoor, Sai .................... 73
Cios, Krzysztof ............... 92  Newell, Karen ................... 50
Clement, Mark ................. 17  Quiogue, Leigh ................ 90
Collins, Nicole ............... 1  Rich, Ivan .......................... 56
Crandall, Keith ............... 17  Richidi, Thami .................. 92
Cushman, John ................. 11  Selvam, Priyadarshini ....... 30
Deshmukh, Ankur ............. 30  Seppi, Kevin ................... 17
Giclas, Patricia ............... 50  Snell, Quinn ................... 17
Gillery, Martin ............... 11  Snow, Neil ........................ 71
Guan, Zhijie ................... 81  Srikantaswamy, Aparna ...... 50
Gunawardane, Ru ............ 1  Tripet, Brian ..................... 73
Habib, Mona Soliman ....... 83  Tye, Virgil Gary ............... 36
Hall, Karen .................... 56  Witt, Abigail .................. 1
Hans, Reena ................... 30  Wolkow, Tom ................... 68
Hines, Lisa ..................... 1  Zukoski, Matthew J .......... 4
Hodges, Robert ............... 73  Zurek, Daniel M ................ 89