ISBRA 2007

INTERNATIONAL SYMPOSIUM ON
BIOINFORMATICS RESEARCH
AND APPLICATIONS

Poster Proceedings

MAY 7-10
Department of Computer Science
Georgia State University
Atlanta, Georgia

http://www.cs.gsu.edu/isbra
Message from the Poster Chairs

On behalf of Program Committee, we would like to welcome you to the 2007 International Symposium on Bioinformatics Research and Applications (ISBRA 2007). The ISBRA symposium is the successor of the International Workshop on Bioinformatics Research and Applications (IWBRA), held on May 22-25, 2005 in Atlanta, GA and on May 28-31, 2006 in Reading, UK, in conjunction with the International Conference on Computational Science.

This year, 27 poster papers were submitted in response to the call for poster papers. Following a review process, the program committee has selected 23 poster papers for presentation at the symposium. The topics include genome analysis, disease association and disease susceptibility prediction, clustering and classification, gene expression analysis, gene networks, motif finding, pathways, protein structure and interaction, phylogenetics, and software tools.

We would like to thank all authors for submitting poster papers and presenting their work at the symposium. We extend special thanks to the Organizing, Publications, Finance, Publicity, and Posters Chairs, all of whom are listed on the following page, for their tremendous efforts in making ISBRA 2007 a great success. Last but not least, we would like to thank the General Chairs, Dan Gusfield and Yi Pan, for their leadership and guidance.

We hope you will find the poster program interesting and thought provoking and that attending ISBRA 2007 will provide you with stimulating ideas and ample opportunities to meet other researchers from around the world. Enjoy poster sessions!

Dumitru Brinza
Ph.D. Candidate
Department of Computer Science
Georgia State University
34 Peachtree, suite 1415
Atlanta, Georgia, 30303, USA
phone: 404-463-2808
e-mail:dima@cs.gsu.edu

Gulsah Altun
Ph.D. Candidate
Department of Computer Science
Georgia State University
34 Peachtree, suite 1417
Atlanta, Georgia, 30303, USA
phone: 404-651-5733
e-mail:gulsah@cs.gsu.edu
ISBRA 2007

Conference Organization

General Chairs
Dan Gusfield, University of California, Davis, USA
Yi Pan, Georgia State University, USA

Program Chairs
Ion Mandoiu, University of Connecticut, USA
Alexander Zelikovsky, Georgia State University, USA

Organizing Chairs
Robert Harrison, Georgia State University, USA
Yanqing Zhang, Georgia State University, USA

Publications Chair
Raj Sunderraman, Georgia State University, USA

Finance Chairs
Anu Bourgeois, Georgia State University, USA
Akshaye Dhawan, Georgia State University, USA

Publicity Chairs
Kim King, Georgia State University, USA
Yingshu Li, Georgia State University, USA

Poster Chairs
Dumitru Brinza, Georgia State University, USA
Gulsah Altun, Georgia State University, USA

Graphic Designer
Stefan Gremalschi, Georgia State University, USA
Poster Session A

A1. Greedy Approach to Reliable Disease Susceptibility Prediction .............................................1
   Dumitru Brinza, Irina Astrovskaya, and Alex Zelikovsky
   One of the main problems in genetic epidemiology is to robustly predict genetic susceptibility to complex
diseases based on the data from case/control studies. This becomes computationally challenging in presence
of interactions between multiple genes. In order to efficiently search through enormous amount of possible
combinations, it is necessary to apply heuristics and the greedy approach has been successfully validated on
many real data. In this paper we modify the genotype covering phase of the model-fitting susceptibility
prediction algorithm from (Brinza and Zelikovsky, WABI06). We improve reliability of the previously known
prediction method based on greedy approach by replacing “clean” genotype coverage with “dirty” coverage, i.e.,
clusters participating in coverage do not overlap or overlap respectively. We have leave-one/many-out cross-
validated existed and proposed prediction methods on real case/control studies of four diseases (Chron’s
disease, autoimmune disorder, tick-born encephalitis, and lung cancer). Our results show that relaxation of the
clean coverage significantly improves reliability of the greedy based susceptibility prediction approach.

A2. Optimal Mapping of Multisource Trees into DAGs in Biological Networks ..............................5
   Qiong Cheng and Alex Zelikovsky
   The Homeomorphism problem of pair wise graphs has been an open problem in the theory of graph. There
have been lots of applications which can be reducible to the problem such as evolution of pathways and
reconstruction of biological networks. We aim to propose a feasible solution to calculate the minimal cost of
homeomorphism alignment in biological pathways. In our first stage, we implement the dynamical program
algorithm to align the directed cyclical and acyclical graphs in pair wise.

A3. Toward a Methodology for Discovery of Regulatory Motifs in Plant ........................................9
   Dazhang Gu, Klaus H. Ecker, Lonnie Welch, and Sarah Wyatt
   Motif discovery is valuable in finding elements regulating the growth of plants, which has significant agricultural
values. However, previous work in evaluation of motif discovery tools and techniques has focused primarily on
prokaryotes and animals. Thus it is not clear how well these tools and techniques work for plant genomes. Most
existing benchmarks are not representative of nucleotide biases found in promoters in plant species. In addition,
there is a lack of methodology for discovery of regulatory motifs. This paper addresses the first two problems.
First, we describe our findings about the characteristics of the currently known regulatory motifs in Arabidopsis
thaliana. Then we describe how this information is used to generate representative benchmarks which are used
to evaluate motif discovery tools.

A4. Discovering Causal Sentences with Automatically Learned Patterns .........................................13
   Shreekanth Karvaje, Bharat Ravisekar, Baoli Li, and Ashwin Ram
   We propose a semi-supervised method to extract rule sentences from medical abstracts. Medical rules are
sentences that give interesting and non-trivial relationship between medical entities. Mining such medical rules
is important since the rules thus extracted can be used as inputs to an expert system or in many more other
ways. The technique we suggest is based on paraphrasing a set of seed sentences and populating a pattern
dictionary of paraphrases of rules. We match the patterns against the new abstract and rank the sentences.

A5. An Imputation Method Based on Robust Regression Using Minimum Covariance Determinant
    Estimates for Cancer Gene Expression Data .................................................................17
   Hyunsoo Kim and Haesun Park
   There sometimes exist outliers in cancer gene expression datasets built from various tumor tissues and/or
different patients, which can adversely affect data analysis results. In this paper, we introduce an imputation
method based on robust regression using minimum covariance determinant (MCD) estimates. In our
experiments, our method achieved good performance for a cancer microarray gene expression dataset that has
outliers.

A6. Design Pattern for Protein Identification .................................................................21
   Jens Lichtenberg and Lonnie Welch
   This paper presents the protein identification problem in form of a design pattern. Christopher Alexander defines
patterns as things that happen in the world and rules, which tell how to create such things and when they must
be created. The pattern presented in this article addresses the identification of the correct protein based on a
given peak list. It provides a process on how to infer the identity and primary structure of such protein based on
A7. Characterizing Pseudobase and Predicting RNA secondary structure with simple H-type pseudoknots

Oyun-Erdene Namsrai

After we had analyzed all the pseudoknots in Pseudobase, we classified it into six basic pseudoknot types. Of 245 unique pseudoknots in PseudoBase database, 235 were the H-type pseudoknots. Our characterization of Pseudobase shows that efficient and effective tools for detecting the RNA H-type pseudoknots are needed. In this paper, we present a computation the maximum number of base pairs of an RNA sequence with simple H-type pseudoknots. Our algorithm called “Iterated Base Pair Maximizing” is based on the base pair maximizing algorithm (Nussinov et al., 1978). To evaluate our method we use the more than 20 sequences with simple H-type pseudoknots of variable size from 19 to 30 nucleotides. We get our experimental data set from PseudoBase, the algorithm correctly identifies more than 75% of base pairs for short sequences and 65% overall. It correctly predicts nearly all H-type pseudoknots. In addition, our approach can be easily extended and applied to other classes of more general pseudoknots. Availability: The algorithm has been implemented in C++ in a program called IBPM, which is available at http://dblab.cbu.ac.kr/ibpm.

A8. Mining MEDLINE for Gene Clustering: A Comparison of Feature Selection Approaches

Sailaja Pydimarri, Orlando Karam, Venu Dasigi, and Rajnish Singh

In this paper we compare two different schemas for feature selection applied to data mining of MEDLINE data. We also highlight the importance of the choice of background set for both of these approaches. We compare the performance of these schemas and the choice of background sets both by keyword evaluation by an expert and by applying it to document clustering.


Gregory Reck and Iosif Vaisman

A statistical potential function has been derived based on Delaunay tessellation of a large number of computationally hydrated protein structures for eventual use in mutant stability studies. The tessellation provides quadruplets of “nearest neighbor” residues and the potential function is derived from the observed occurrence of each quadruplet compared to a reference expectation. Several potential functions are computed from the hydrated structures using alternative reference expectations and compared with a function from unhydrated structures using a random distribution expectation. Each of the potential functions is used to score groups of single point mutants of barnase, staphylococcal nuclease and T4 lysozyme, and the scores are correlated with reported stability measurements of the mutant proteins. Three machine learning tools (decision tree, random forests, support vector machine) are used to demonstrate a relationship between the potential function scores and the stabilities of the mutants.

A10. A Topological Characterization of Protein-Water Interactions for Knowledge-Based Models

Gregory Reck and Iosif Vaisman

A new strategy is described that characterizes the relationship between a globular protein and its water environment. This technique uses the 4-body nearest-neighbor relationships defined by Delaunay tessellation of computationally hydrated proteins. The new water group parameter is based on the extent of inclusion of simulated water molecules into the Delaunay tetrahedra surrounding each residue. Topological parameters are especially informative in developing knowledge-based methods for predicting protein structure and correlating protein behavior such as stability and functionality.

A11. Integrated Statistical and Association Rule Analyses of Time-Dependent Gene Co-expression Patterns

Sandra Rodriguez-Zas, Younhee Ko, and Bruce Southey

An integrated approach that combines powerful statistical and data mining tools with functional annotation information was used to fully characterize gene expression data. First, filtering, normalization and linear modeling of gene expression observations are used to minimize the impact of technical or nuisance variation on gene expression intensities and identify genes with differential expression across conditions. Then, condition-dependent thresholds are applied to the adjusted gene expression patterns and quantitative association rules are obtained. Interesting relationships among genes are identified based on the simultaneous consideration of different quality rule criteria. The integrated approach exploited the advantages of hierarchical models and association rule, assisting in the interpretation of gene expression data and providing novel insights into several sociogenomic pathways.
Poster Session B

B1. **CABIN: Collective Analysis of Biological Interaction Network** ........................................ 45
   Mudita Singhal and Kelly Domico
   CABIN is developed as a plugin to Cytoscape (Shannon et al.) - an open source network visualization and analysis tool. CABIN promotes analytical reasoning for integrating evidence of interaction data from multiple sources by the use of interactive visual interfaces. Multiple coordinated views within CABIN fosters exploratory data analysis by users accommodating for expert domain knowledge. The functionalities available within CABIN maximize human perception and understanding of uncertain and complex data facilitating high quality human judgment with limited investment of the user's time.

B2. **Smoothing spline mixed effects modeling of multi-factorial gene expression profiles** ................... 47
   Brandon Smith, Bruce Southey, and Sandra Rodriguez-Zas
   Time-course gene expression data from microarray studies provide unique insights into the genes and pathways. The analysis of time-course microarray experiments is challenging because of the wide range of gene expression patterns, multiple sources of variation, and structure of the time-dependent measurements. A smoothing spline mixed model is proposed to describe complex time-course gene expression. A penalized likelihood-based approach was used to obtain estimates. The flexible approach was used to characterize the patterns of gene expression during honey bee behavioral maturation in a multifactorial experiment including time, race and host colony effect. Two spline dimensional bases were considered and a likelihood-based criterion was used to select a parsimonious spline basis that accommodates all the trajectories observed. Gene Ontology information of the microarray elements with significant differential expression across time within and across races and hosts indicated that, among others, genes that regulate development and synaptic connections are associated with maturation.

B3. **DeltaProt: Molecular comparison of proteins based on sequence alignments** .......................... 51
   Steinar Thorvaldsen, Tor Fle, and Nils P Willassen
   In the toolbox DeltaProt we present statistical methods and trend-tests that are useful when the protein sequences in alignments can be divided into two or more groups based on known phenotypic traits such as preference of temperature, pH, salt concentration or pressure. The algorithms have been successfully applied in the research on extremophile organisms. We also provide procedures to plot the output from these tests for visualisations. DeltaProt is a Matlab© companion Toolbox that can be used freely for academic, non-profit purposes. Available from [http://www.math.uit.no/bi/deltaprot/](http://www.math.uit.no/bi/deltaprot/).

B4. **Using Logical Sets to Target Gene Expression Patterns** ..................................................... 55
   Timothy Tickle and M. Taghi Mostafavi
   A quantitative method of determining significant genes from microarray data based on their experiment-wide expression pattern has been developed. This technique involves combining different sets of comparisons with different criteria for their comparisons on p-values derived from two-sided tests to target different gene expression patterns. This technique is able to target specific gene expression patterns in clinical micro array data based on ovarian cancer tissue samples (primary tumors). One result from a previous pilot study which is also found as a result in this analysis was confirmed with real-time RT-PCR. Other genes found have known associations with cancer.

B5. **Database for Structural Analysis of HIV Protease** ............................................................ 59
   Yunfeng Tie, Hao Wang, Robert Harrison, and Irene Weber
   A prototype domain-specific database has been designed to organize and automate the comparative analysis of protein structures. The database is designed to provide the basic functions of structural analysis with a flexible modular computing architecture. Structure data are represented by a hierarchical arrangement of objects for atom, residue, chain and PDB. The current implementation includes functions for superimposing up to 8 structures on a reference structure, and calculation of user-selected interatomic distances. This database has been tested with crystal structures of HIV protease mutants in complex with various inhibitors. Antiviral HIV protease inhibitors developed with the aid of structures are important drugs to treat HIV/AIDS. This database for comparison of structures has the potential for further development for improved analysis of HIV protease mutant structures, drug design, and extension to structural analysis of other proteins.
B11.

A survey of basecalling algorithms .............................................................. 68
Andy Ju An Wang

Basecalling is critical for human genetics and bioinformatics research. The term, definition, and process of basecalling have been evolving over time with a variety of standards and technologies coexisted in the literature. It can be difficult for beginners to navigate and understand the most important aspects of basecalling and associated algorithms. This paper provides a brief overview of basecalling algorithms for computer scientists who are interested in bioinformatics research and development. The major basecalling algorithms discussed here include ABACUS, Affymetrix GDAS, and Model-P.

B12. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.

B13. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.

B14. Weighbor, BIONJ and NJ programs. Besides, this poster presents validation of proposed solutions.

B15. Directed a matrix to predict residue pairing based upon high mutual information. Residues with high mutual information are statistically closer to each other in tertiary structure other than random expectations. This characteristic for co-evolving residues provide the sequence information for MI analysis. Residues with high mutual information are statistically closer to each other in tertiary structure other than random expectations. This characteristic for co-evolving residues could be used as an additional parameter to help predict protein-protein interaction sites. We analyzed the residue pairing preferences with all the possible co-evolving residues from these 48 complex pairs and generated a matrix to predict residue pairing based upon high mutual information.

B16. Co-evolution analysis of protein complexes and its applications in pairing preferences prediction ....81
FangFang Pan, Dongsheng Che, Michelle Momany, Liming Cai, and Russell Malmberg

Co-evolution between protein interacting partners is a topic of great interest. The dramatic increase in the number of available protein sequences and structures allows the application of mutual information to protein co-evolution analysis. We found 48 hetero interaction pairs in protein complexes with known 3D structures to provide the sequence information for MI analysis. Residues with high mutual information are statistically closer to each other in tertiary structure other than random expectations. This characteristic for co-evolving residues could be used as an additional parameter to help predict protein-protein interaction sites. We analyzed the residue pairing preferences with all the possible co-evolving residues from these 48 complex pairs and generated a matrix to predict residue pairing based upon high mutual information.

B17. Weighbor, BIONJ and NJ programs. Besides, this poster presents validation of proposed solutions.

B18. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.

B19. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.

B20. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.

B21. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.
Greedy Approach to Reliable Disease Susceptibility Prediction

Dumitru Brinza*, Irina Astrovskaya, and Alexander Zelikovsky**

Department of Computer Science, Georgia State University, Atlanta, GA 30303
{dima, iraa, alexz}@cs.gsu.edu

Abstract. One of the main problems in genetic epidemiology is to robustly predict genetic susceptibility to complex diseases based on the data from case/control studies. This becomes computationally challenging in presence of interactions between multiple genes. In order to efficiently search through enormous amount of possible combinations, it is necessary to apply heuristics and the greedy approach has been successfully validated on many real data. In this paper we modify the genotype covering phase of the model-fitting susceptibility prediction algorithm from [1]. We improve reliability of the previously known prediction method based on greedy approach by replacing clean genotype coverage with dirty coverage, i.e., clusters participating in coverage do not overlap or overlap respectively. We have leave-one/many-out cross-validated existed and proposed prediction methods on real case/control studies of four diseases (Chron’s disease, autoimmune disorder, tick-born encephalitis, and lung cancer). Our results show that relaxation of the clean coverage significantly improves reliability of the greedy based susceptibility prediction approach.

1 Introduction

The difference between individual DNA sequences occurs at a single-base sites, in which more than one allele is observed across population. Such variations are called single nucleotide polymorphisms (SNPs). The number of simultaneously typed SNPs for association and linkage studies is reaching 10^6 for SNP Mapping Arrays. Some complex diseases, such as psychiatric disorders, are characterized by a non mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other. In general, a single SNP or gene may be impossible to associate because a disease may be caused by completely different modifications of alternative pathways. Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease. It is even difficult to decide if a particular disease is genetic, e.g., the nature of Crohn’s disease has been disputed. Although answers to above questions may not explicitly help to find specific disease-associated SNPs, they

* Supported by GSU Molecular Basis of Disease Fellowship
** Partially supported by NSF Award 0429735 and US CRDF Award MOM23049CS03
may be critical for disease prevention. Indeed, knowing that an individual is (or is not) susceptible to (or belong to a risk group for) a certain disease will allow greatly reduce the cost of screening and preventive measures or even help to completely avoid disease development, e.g., by changing a diet.

The problem addressed in this work is the disease susceptibility prediction problem exploiting the developed methods for searching associated risk and resistance factors. We improve reliability of the previously known prediction method based on greedy approach by replacing clean genotype coverage with dirty coverage, i.e., clusters participating in coverage do not overlap or overlap respectively. We have leave-one/many-out cross-validated existed and proposed prediction methods on real case/control studies of five diseases. Our results show that relaxation of the clean coverage significantly improves reliability of the greedy based susceptibility prediction approach.

2 Heuristics for Disease Susceptibility Prediction

The following is the formal description of the problem from [1].

Disease Susceptibility Prediction Problem. Given a sample population $S$ (a training set) and one more individual $t \notin S$ with the known SNPs but unknown disease status (testing individual), find (predict) the unknown disease status.

Ideally, there should be two clusters perfectly distinguishing diseased from control individuals. There is a trade-off between number of clusters and the information carried by clustering which results in trade-off between number of errors (i.e., incorrectly clustered individuals) and informativeness which was proposed to measure by information entropy instead of number of clusters [1].

Optimum Disease Clustering Problem. Given a population sample $S$, find a partition $\mathcal{P}$ of $S$ into clusters $S = S_1 \cup \ldots \cup S_k$, with disease status 0 or 1 assigned to each cluster $S_i$, minimizing

$$\text{entropy}(\mathcal{P}) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} \ln \frac{|S_i|}{|S|}$$

for a given bound on the number of individuals who are assigned incorrect status in clusters of the partition $\mathcal{P}$, $\text{error}(\mathcal{P}) < \alpha \cdot |\mathcal{P}|$.

The above optimization formulation is obviously NP-hard but has a huge advantage over the prediction formulation that it does not rely on cross-validation and can be studied with combinatorial optimization techniques. Still, in order to make the resulted clustering algorithm useful, one needs to find a way ho to apply it to the original prediction problem.

The following general approach has been proposed by us [1]. Assuming that the clustering algorithm indeed distinguishes real causes of the disease, one may expect that the major reason for erroneous status assignment is in biases and lack of sampling. Then a plausible assumption is that a larger sample would lead to a
lesser proportion of clustering errors. This implies the following transformation of clustering algorithm into prediction algorithm:

**Clustering-based Model-Fitting Prediction Algorithm**

- Set disease status 0 for the testing individual $t$ and find the optimum (or approximate) clustering $P_0$ of $S \cup \{t\}$
- Set disease status 1 for the testing individual $t$ and find the optimum (or approximate) clustering $P_1$ of $S \cup \{t\}$
- Find which of two clusterings $P_0$ or $P_1$ better fits model, and accordingly predict status of $t$,

$$\text{status}(t) = \arg\min_{i=0,1} \text{error}(P_i)$$

Two clustering algorithms based the combinatorial and complementary greedy association searches has been proposed in [1]. This clustering finds for each individual an MSC or its cluster that contains it and is the most associated according to a certain characteristic (e.g., RR, PPV or lowest p-value) with disease-susceptibility and disease-resistance. Then each individual is attributed the ratio between these two characteristic values – maximum disease-susceptibility and disease-resistance. Although the resulted partition of the training set $S$ is easy to find, it is still necessary to decide which threshold between case and control clusters should be used. The threshold can be chosen to minimize the clustering error.

The **complimentary greedy search-based prediction algorithm** (CGSP) exploits complimentary greedy search to find the most-associated cluster for each individual. Empirically, the best association characteristic is found to be the positive predictive value PPV. In this paper we modified the *clean* clustering (i.e., clusters do not overlap) introduced in [1] to the *dirty* clustering (i.e., clusters might overlap). The leave-one-out and leave-many-out cross-validation (see Section 3) show some advantages of *dirty* GCSP over previously published *clean* GCSP prediction algorithm for considered real datasets.

### 3 Experimental results

#### Datasets

- **Crohn’s disease** – dataset [4] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn’s disease by genotyping 103 SNPs for 129 trios. All offspring belong to the case population, while almost all parents belong to the control population. In entire data, there are 144 case and 243 control individuals.

- **Autoimmune disorder** – dataset [6] is sequenced from 330kb of human DNA containing gene CD28, CTLA4 and ICOS which are proved related to autoimmune disorder. A total of 108 SNPs were genotyped in 384 cases of autoimmune disorder and 652 controls.

- **Tick-borne encephalitis** – dataset [3] consists of 41 SNPs genotyped from DNA of 21 patients with severe tick-borne encephalitis virus-induced disease and 54 patients with mild disease.
Table 1. Leave-One-Out (LOO) and Leave-Half-Out (LHO) cross-validation of 4 disease-association search methods on 4 real datasets.

<table>
<thead>
<tr>
<th>Data</th>
<th>Cross validation</th>
<th>Prediction Method</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chron’s disease</td>
<td>LOO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>93</td>
<td>47</td>
<td>196</td>
<td>51</td>
<td>74.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>79</td>
<td>31</td>
<td>212</td>
<td>65</td>
<td>75.2</td>
</tr>
<tr>
<td></td>
<td>LHO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>61</td>
<td>58</td>
<td>185</td>
<td>79</td>
<td>64.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>73</td>
<td>47</td>
<td>196</td>
<td>71</td>
<td>69.5</td>
</tr>
<tr>
<td>Tick-borne encephalitis</td>
<td>LOO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>19</td>
<td>5</td>
<td>60</td>
<td>7</td>
<td>86.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>23</td>
<td>3</td>
<td>62</td>
<td>3</td>
<td>93.4</td>
</tr>
<tr>
<td></td>
<td>LHO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>1.92</td>
<td>8.79</td>
<td>38.31</td>
<td>7.61</td>
<td>71.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>5.38</td>
<td>9.94</td>
<td>37.61</td>
<td>4.12</td>
<td>75.4</td>
</tr>
<tr>
<td>Lung cancer (Germans)</td>
<td>LOO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>210</td>
<td>0</td>
<td>228</td>
<td>44</td>
<td>90.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>196</td>
<td>0</td>
<td>228</td>
<td>58</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>LHO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>35.07</td>
<td>48.14</td>
<td>159.32</td>
<td>59.68</td>
<td>64.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>6.1</td>
<td>12.85</td>
<td>197.5</td>
<td>87.15</td>
<td>67.0</td>
</tr>
<tr>
<td>Autoimmune disorder</td>
<td>LOO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>106</td>
<td>31</td>
<td>621</td>
<td>278</td>
<td>70.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>97</td>
<td>29</td>
<td>623</td>
<td>287</td>
<td>69.5</td>
</tr>
<tr>
<td>Autoimmune disorder</td>
<td>LHO</td>
<td>CGSP\textit{Dirty}−cover</td>
<td>133</td>
<td>126</td>
<td>526</td>
<td>251</td>
<td>63.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGSP\textit{Clean}−cover</td>
<td>131</td>
<td>88</td>
<td>564</td>
<td>253</td>
<td>67.1</td>
</tr>
</tbody>
</table>

Lung cancer – dataset [5] was obtained using genome-wide DNA pooling strategy in a group of German smokers clinically diagnosed with lung cancer and age-matched healthy smokers. 83,715 SNPs had been screened. 141 SNPs showed putative allelic imbalance between case and control DNA pools and were eventually genotyped in individual samples. Finally, dataset includes 322 male smokers with lung cancer and 273 healthy smokers genotyped in 141 SNPs.

References

Optimal Mapping of Multi Source Trees into DAG in Biological Network

Qiong Cheng, Alexander Zelikovsky

Department of Computer Science, Georgia State University, Atlanta, GA 30303, {cscqcxz,alex}@cs.gsu.edu

Abstract. The effective and efficient prediction and reconstruction of biological networks for a variety of inter-species and intra-species organisms have been a challenging research topic in computational system biology. They are becoming more and more important with the immense increase of good-quality data now available from high-throughput genomic, proteomic and metabolomic sources.

Our research focuses on the network alignment problem on a class of biological networks which can be modeled as directed acyclic graphs (DAG) such as phylogenetic networks, and metabolic pathways. Our objective is to obtain the minimal cost of mapping of multi source tree into DAG; and our approach is to employ dynamic programming to obtain the approximately optimal solution. We implemented the algorithm and are evaluating the preliminary results on metabolic network.

Further efforts are underway to analyze the computational complexity of the specified DAG mapping problem and to build a general tool for automatically mining feasible patterns for the reconstruction of multiple bimolecular networks at different levels and dimensions.

Key words: System Biology, Graph Theory, Graph Mapping

1 Introduction and Related Work

Graph theoretic formalism is adopted to study the subgraph mapping problem. Some papers in pathway mapping only consider the maximal similarity of sequences composing the network. Their authors assumed that the networks have the same fixed topology. However there are some others who consider not only sequence similarity but also network topological similarity. Pinter et al. modeled metabolic network of pathways as outgoing trees with different size and reduced the problem to the approximate labeled tree homeomorphism one[1]. Kelly et al. rebuilt a global alignment graph in which each vertex represents a pair of proteins and each edge represents a conserved interaction, gap, or mismatch; and then they try to find the highest-scoring path with limited length and no consecutive gaps or mismatches([4], [5] and [6]).

Both vertex similarity and network topology similarity are considered in network alignments. Chen et al. formulated the problem as an integer quadratic
problem to obtain the global similarity score based on the mapping of as many as node-node similarity and as many as edge-edge similarity[2]. Yang et al. employed an exhaustive searching approach to find the vertex-to-vertex and path-to-path mappings having the maximal mapping score with the limited length of gaps or mismatch[3].

To explore the general mapping problem, we focus on the solution to a class of problems of mapping of multi source tree to the networks which can be modeled as directed acyclic graphs. A solution based on dynamic programming is proposed in this paper so that a vertex-to-vertex and edge-to-path mappings with the minimal homomorphism cost is obtained. The flexible interface can be provided to support different vertex similarity calculation and different approaches to generate random graphs.

2 Network Alignment Problem Formulation

2.1 NOTATION AND DEFINITIONS

Definition 1. A directed path is a sequence of distinct vertices \( v_0, v_1, ..., v_i, ..., v_k \) and edges \( e_0, e_1, ..., e_i, ..., e_{k-1} \) such that \( e_i \) is an edge directed from \( v_i \) to \( v_{i+1} \), for all \( i < k \). We denote the path with the first vertex \( v_0 \) and the last vertex \( v_k \) by \( P_{0,k} \). In this paper, path will always mean 'directed path’. Two or more disjoint Paths mean that they share at least no last vertex with each other.

Definition 2. Pattern tree \( T = < V_T, E_T, L_T > \) is a multi-source tree in which \( V_T \) is the vertex set, \( E_T \) is the edge set, and \( L_T \) is the label set. Every vertex in \( V_T \) has unique label in \( L_T \). The multi-source tree is a directed acyclic graph (DAG), whose underlying undirected graph is a tree [1].

Definition 3. Text \( G = < V_G, E_G, L_G > \) is a directed acyclic graph (DAG) in which \( V_G \) is denoted as vertex set, \( E_G \) is denoted as edge set, and \( L_G \) is denoted as label set. Every vertex in \( V_G \) has its unique label in \( L_G \).

Definition 4. A mapping from pattern \( T = < V_T, E_T, L_T > \) to text \( G = < V_G, E_G, L_G > \) is called homomorphic if

1. every vertex in \( V_T \) is mapped to a vertex in \( V_G \) \( (f_v : V_T \rightarrow V_G) \);
2. every edge \( e = (u, v) \) in \( E_T \) is mapped to a path in \( P_G \) \( (f_e : E_T \rightarrow P_G) \). The path is an ordered sequence \( u_0 = f(u), u_1, u_2, ..., u_k = f(v) \) of vertices in \( P_G \);

Definition 5. Let \( \Delta : L_T \times L_G \rightarrow \mathbb{R} \) denote the 2-place relation on the set of all ordered pairs of a vertex in \( V_T \) and a vertex in \( V_G \) to a real number. \( \delta \) represents the penalty score for deleting a vertex from a path. The cost of a homomorphic mapping \( f : T \rightarrow G \) is

\[
\text{cost}(f) \sum_{v \in V_T} \Delta(v, f(v)) + \delta \times \sum_{e \in E_T} (|f(e)| - 1)
\]
2.2 PROBLEM FORMULATION

Given pattern tree $T$ and text graph $G$, find a homomorphic mapping that minimize the homomorphic cost. The homomorphic mapping allows vertices repeated occurrence.

An example of the mapping between metabolic pathways are as follows. Our objective in the example is to find a mapping of pattern graph into the pyrimidine ribonucleotide ribonucleoside metabolism pathway so that the mapping score is minimized.

![Pattern Graph and Text Graph](image)

Fig. 1. Pattern graph is a multi source tree. Text graph is a pyrimidine ribonucleotide ribonucleoside metabolism pathway which is a DAG.

2.3 Dynamic Program Algorithm

We build a two dimension table $DT$ for our dynamic program by the post order traversal. Every column is a homomorphic mapping from every vertex of pattern $T$ to a vertex of text $G$. Every row is a mapping from a vertex of pattern $T$ to every vertex of text $G$. The default value in every entry is initialized to infinity.

The recursive function for the dynamic program is as follows,

$$DT(u^T, v^G) = \min\{\Delta(u^T, v^G) + \sum_{u^T_j \in u^T.children} \min_{v^G_i \in v^G.children} DT(u^T_j, v^G_i), \quad (1)\}$$

$$PenaltyOfDelete(v) + \min_{v^G_i \in v^G.children} DT(u, v^G_i)\} \quad (2)$$

$$PenaltyOfDelete(v) + \min_{v^G_i \in v^G.children} DT(u, v^G_i)\} \quad (2)$$
3 Future Work

More meaningful experiments are in need. We will generate the pattern graph randomly by network motifs with small number of vertices of different and random labels such as with the size of 3, 4 and 5 vertices[7]. These generated patterns will be used to verify the reliability of our algorithms.

4 Conclusion

We devote ourselves to the studies of mining graph data and their application of studying biological evolution and reconstruction. We reformulated the problem proposed in [1], designed and implemented the dynamic programming. Further efforts are in need to evaluate our algorithm so that we could have a deeper understanding towards the studies of biological evolution and reconstruction.

Acknowledgments. We acknowledge Dr. Ron Y Pinter and his student Oleg Rokhlenko for sharing their executable code and graphic user interface with us.

References

2. Zhenping Li, Yong Wang, Shihua Zhang, Xiang-Sun Zhang, Luonan Chen, Alignment of Protein Interaction Networks by Integer Quadratic Programming EMBS '06. 28th Annual International Conference of the IEEE : 5527-5530, (Aug. 2006)
Toward a Methodology for Discovery of Regulatory Motifs in Plants

Dazhang Gu\textsuperscript{1}, Lonnie Welch\textsuperscript{1,2,3}, Sarah Wyatt\textsuperscript{2,4}, and Klaus Ecker\textsuperscript{1}

\textsuperscript{1} School of Electrical Engineering and Computer Science  
\textsuperscript{2} Molecular and Cellular Biology Program  
\textsuperscript{3} Biomedical Engineering Program  
\textsuperscript{4} Department of Environmental and Plant Biology  
Ohio University, Athens, Ohio 45701  
\{gud,welch,wyatts,ecker\}@ohio.edu

Motif discovery is valuable for finding genomic elements that regulate the growth of plants. Such knowledge has important implications in areas such as agriculture and medicine. However, previous work in the evaluation of motif discovery tools and techniques has focused primarily on prokaryotes and animals [1, 2]. Most existing motif discovery benchmarks are not representative of nucleotide biases found in promoters in plant species. In addition, there is a lack of methodology for discovery of regulatory motifs. Since genomic properties of plants differ significantly from other organisms, it is not clear how effective the tools and techniques will be for plant genomes. In this paper we describe our findings about the characteristics of the currently known regulatory motifs in the model plant \textit{Arabidopsis thaliana}. Then we describe how this information is used to generate representative benchmarks which are used to evaluate the effectiveness of motif discovery tools for plants.

\textbf{Fig. 1.} Histogram of Known Motifs in Arabidopsis (Agris)
We begin by examining the characteristics of the motifs in plants. Arabidopsis thaliana is a well researched plant in biology, and many of its transcription factor binding site motifs have been identified. The features of the motifs offer us useful insights when analyzing plant genomes. An extensive list of Arabidopsis motifs can be found on the Agris website [3]. In our study, the motifs are examined in terms of the motif length and the number of variations. The histogram in Figure 1 reveals their distribution. The X-axis of the graph represents the total length of a motif, including variable letters; the Y-axis represents the number of variable letters in a motif; the Z-axis represents the count of known motifs at the given motif length and number of variations. It is interesting to see that the majority of known binding site motifs have 10 or fewer nucleotides and contain zero or one variable letter.

In order to evaluate and compare the strengths of different motif discovery tools for motifs of different lengths and variations in Arabidopsis, 30 data sets were generated for 30 representative motifs selected from the known motifs; the data sets were input to the motif discovery tools. Each data set contains 5 to 20 sequences randomly, which were generated following a distribution that is like the distribution of nucleotides in Arabidopsis promoters. The length of each sequence varies randomly from 1000 to 3000, and the motif of a data set is embedded in each of the sequences in the data set at a random location.

Using these data sets, we evaluated six motif discovery tools: MDSan (MD) [4], AlignACE (ACE) [5], MEME (MEME) [6], Teiresias (Tei) [7], MotifSampler (MSam) [8], and BioProspector (BioP) [9]. Since the embedded motifs are known, sensitivity and specificity are measured for the motifs predicted by each tool. Additionally, a combined metric, F-score, is computed [2]. To compare the power of different tools for various kinds of motifs, these 30 data sets are further grouped into 4 main categories (quadrants) that are partitioned into two clusters based on length, and into two based on variation. Based on the information from Figure 1, the lengths are chosen to be short (<10) or long (=10), and the variations are chosen to be low (<2) or high (=2). The motif discovery tools are evaluated and compared based on F-scores in each of the quadrants. The results are shown in Figure 2. Levels of variation are shown along the X-axis and lengths are shown along the Y-axis. In each graph, the X-axis represents the various tools and the Y-axis represents the F-score. The wide gray bar stands for the mean value of the samples. It can be seen that when motifs are short and have high variation, the discovery is more difficult since the motifs are easily covered by background noise. MEME and Teiresias perform quite well for long motifs with low variation; MEME and AlignACE perform best for long motifs with high variation. The comparisons are more clearly visualized by computing the confidence intervals (0.95) shown in Figure 3. This information will help plant biologists to make informed choices of motif discovery tools.

In conclusion, our efforts in the areas of evaluation of motif discovery tools in the context of plant genomes constitute important intermediate steps toward the ultimate goal of defining a methodology for the discovery of regulatory motifs.
Fig. 2. F-scores of 6 Motif Discovery Tools

Fig. 3. Confidence Intervals of the F-scores
References


[3] Website: http://arabidopsis.med.ohio-state.edu/AtcisDB/bindingSiteContent.jsp


Discovering Causal Sentences with Automatically Learned Patterns

Shreekanth Karvaje¹, Bharat Ravisekar¹, Saurav Sahay¹, Baoli Li¹, Ernest Garcia², Ashwin Ram¹

¹ College of Computing, Georgia Institute of Technology
{shreek, rbharat, baoli, ashwin}@cc.gatech.edu

² Department of Radiology, Emory University
Atlanta, GA 30322, USA
ernest.garcia@emoryhealthcare.org

Abstract. We propose a semi-supervised method to extract rule sentences from medical abstracts. Medical rules are sentences that give interesting and non-trivial relationship between medical entities. Mining such medical rules is important since the rules thus extracted can be used as inputs to an expert system or in many more other ways. The technique we suggest is based on paraphrasing a set of seed sentences and populating a pattern dictionary of paraphrases of rules. We match the patterns against the new abstract and rank the sentences.

Introduction

Huge amounts of medical rules are scattered in unstructured format in medical publications, journal articles and web resources. Manually extracting these rules is often a labor-intensive and a time-consuming endeavor. In this paper, we propose a very flexible semi-supervised method based on bootstrapping, to extract such knowledge rules. These extracted rules can be fed to an expert system like the PERFEX [1].

We consider medical knowledge as being present in sentences containing a notion of causality. In English, the causative constructions can be explicit, semi-explicit or implicit. Explicit causative constructions contain unambiguous and relevant keywords such as cause, effect, consequence that define the type of relation. The concerned entities are further linked either using causal verbs or causative links (complex clauses and phrases). In this paper we are considering explicit and semi-explicit causative verbs as to be representing causality.
The underlying assumption we make is that if a sentence is known to be containing a useful rule, then sentences that are its paraphrases might contain some rules as well.

Method

Our method of extracting knowledge sentences is as follows:

Building triples database for paraphrasing

A large corpus (like OHSUMED) is used to build a “triples” table which is built according to the method suggested in DIRT [2]. The table contains triplets of path and slots (SlotX and SlotY); and their associated fillers, frequencies and mutual information. We use the Minipar [3] dependency parser to build a dependency tree of each sentence in the corpus and extract paths from this tree using the following criteria:

1. The path should join only content words (Nouns, Verbs, Adjectives, and Adverbs).
2. The paths start and end with entities, in the form of Nouns. This ensures that the resulting patterns link only entities.

To calculate the mutual information between a path p and its filler w, we use the Equation 1, where Slot is either SlotX or SlotY.

\[
mi(p, \text{Slot}, w) = \log \frac{P(p, \text{Slot}, w)}{P(\text{Slot})P(p | \text{Slot})P(w | \text{Slot})}
\]  

Building the pattern dictionary

A small set of seed sentences which represent the knowledge in medical abstracts is input to the system in this step. These sentences are used to build a pattern dictionary. The quality of the patterns extracted depends on the seed rules.

Paths (base patterns) are extracted from the seed sentences as done in the previous step. For each path thus extracted, a matching path is selected from the Triples database, provided one exists. If there is no
direct match, we expand the base pattern based on the nouns that are ending it on either side.

For each base pattern, we find a set of paraphrases using the triples table. These paraphrases form the candidate patterns. The paraphrases are extracted using a similarity score based on the similarity of slots. To calculate the similarity between any two slots slot1 = (p1,s) and slot2 = (p2,s) belonging to two paths p1 and p2, we use the following metric:

\[
Sim(slot_1, slot_2) = \frac{\sum_{w \in T(p_1) \cap T(p_2)} mi(p_1, s, w) + mi(p_2, s, w)}{\sum_{w \in T(p_1)} mi(p_1, s, w) + \sum_{w \in T(p_2)} mi(p_2, s, w)}
\]  

(2)

To calculate the similarity between paths, we propose Equation 3.

\[
Similarity(p_1, p_2) = Max\left(\frac{1}{2}(Sim(p_1, SlotX, p_2, SlotX) + Sim(p_1, SlotY, p_2, SlotY)), \frac{1}{2}(Sim(p_1, SlotX, p_2, SlotX) + Sim(p_1, SlotY, p_2, SlotY))\right)
\]

(3)

Only those candidate patterns that clear a certain thresholds \(\psi, \theta\) (set empirically) for mutual information and similarity respectively are considered.

Selecting and Ranking candidate sentences in new abstract

The algorithm for this step is outlined below:

1. For each sentence in a new document, generate the dependency tree using Minipar and extract paths as done earlier.
2. Each such path extracted, use the triples database to expand this path (by using the similarity, mutual information of the slots and picking the most similar paths).
3. Assign usefulness score to the input sentence using the number of patterns that match the expanded paths. All the sentences in the document that cross a certain threshold \(\theta\) are output.

Evaluations

We evaluated our proposed method on the OHSUMED corpus. OHSUMED collection contains a set of 348,566 abstracts from
MEDLINE. We obtained 127,454 unique paths, 199,844 unique SlotX fillers and 253,541 unique SlotY fillers from the training set.

A set of 38 patterns was used to seed the pattern dictionary. A list of 4,657 paraphrased patterns was extracted. We judged the quality of the patterns by manually evaluating the quality of top 100 patterns. A pattern was adjudged useful if majority of the evaluators have marked it as useful. From this experiment it was clear that on average 75% patterns extracted in the dictionary represent useful patterns.

From the test set, 300 abstracts were manually annotated for rules. Table 1 gives the precision and recall for varying values of threshold $\theta$.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$P$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13.86</td>
<td>90.41</td>
</tr>
<tr>
<td>0.01</td>
<td>13.97</td>
<td>78.08</td>
</tr>
<tr>
<td>0.02</td>
<td>12.84</td>
<td>57.53</td>
</tr>
<tr>
<td>0.05</td>
<td>14.78</td>
<td>46.57</td>
</tr>
</tbody>
</table>

**Conclusion and Future Work**

We have shown that we can extract medical rule sentences with a semi-supervised technique. Our results are quite promising and uphold our assumption that if a sentence is known to be containing a useful rule, then sentences that are its paraphrases might contain some rules as well. In future, we plan to improve the system by tuning the ranking functions used at various levels in our system. Ontologies like WordNet, UMLS could also be incorporated for better accuracy.

**References**

An Imputation Method Based on Robust Regression Using Minimum Covariance Determinant Estimates for Cancer Gene Expression Data

Hyunsoo Kim and Haesun Park
College of Computing, Georgia Institute of Technology,
266 Ferst Drive, Atlanta, GA 30332-0765
{hskim, hpark}@cc.gatech.edu

Abstract. There sometimes exist outliers in cancer gene expression datasets built from various tumor tissues and/or different patients, which can adversely affect data analysis results. In this paper, we introduce an imputation method based on robust regression using minimum covariance determinant (MCD) estimates. In our experiments, our method achieved good performance for a cancer microarray gene expression dataset that has outliers.

Key words: Gene Expression Data Analysis, Missing Value Estimation, Robust Regression, MCD Estimator, Cancer

1 Introduction
Microarray technology is one of the most recent and experimental breakthroughs in molecular biology and it has been applied in a number of interesting studies over a broad range of biomedical problems. Cancer gene experimental data often contain outliers. One of effective methods that can deal with outliers is robust regression using minimum covariance determinant (MCD) estimates [1]. LLSimpute [2] estimates missing values for each gene by grouping the k most similar genes and solving a least squares problem. However, least squares optimization can be adversely affected by outliers. In this paper, we introduce an imputation method based on robust regression using minimum covariance determinant estimates. This method is referred to as LLSimpute/MCD. The rest of this paper is organized as follows. In Section 2, we describe LLSimpute/MCD. Section 3 presents experimental results illustrating the usefulness of our method when there are outliers in a dataset. Summary is given in Section 4.

2 LLSimpute using MCD Estimates
For brevity, we describe LLSimpute/MCD by an example. As an initialization step, all missing values are initially imputed by row averages. For each gene that
Let us assume that a target gene $g_2$ has two missing values in the first and the $n$-th positions. If the missing values are to be estimated by the $k$ most similar genes $(g_{s_1}, 1 \leq i \leq k)$, each element of two matrices $A = [a_{i,j}] \in \mathbb{R}^{k \times (n-2)}$ and $B = [b_{i,t}] \in \mathbb{R}^{k \times 2}$, and a column vector $w = (w_1, \ldots, w_{n-2})^T$ are constructed as

$$
\begin{pmatrix}
g_1^T \\
\cdots \\
g_k^T
\end{pmatrix} =
\begin{pmatrix}
\alpha_1 & w_1 & \cdots & w_{n-2} & \alpha_2 \\
 b_{1,1} & a_{1,1} & \cdots & a_{1,n-2} & b_{1,2} \\
 \vdots & \vdots & \ddots & \vdots & \vdots \\
 b_{k,1} & a_{k,1} & \cdots & a_{k,n-2} & b_{k,2}
\end{pmatrix},
$$

where $\alpha_l (1 \leq l \leq 2)$ is the $l$-th missing value, $w_j (1 \leq j \leq n-2)$ is the $j$-th non-missing element in $g$, $b_{i,t} (1 \leq i \leq k)$ is an element of $g_{s_i}$ in the $l$-th missing position of $g$, and $a_{i,j}$ is an element of $g_{s_i}$ in the $j$-th non-missing position of $g$. For estimating these two missing values, the following least squares formulation can be used:

$$
\min_{\hat{Y}} \| [e_{k \times 1} \ A] \hat{Y} - B \|_F, \tag{1}
$$

where $e_{k \times 1}$ is a $k \times 1$ vector whose all elements are 1,

$$
\hat{Y} =
\begin{pmatrix}
y_{0.1} & y_{0.2} \\
y_{1.1} & y_{1.2} \\
 \vdots \\
y_{n-2.1} & y_{n-2.2}
\end{pmatrix}
= 
\begin{pmatrix}
y_{0.1} & y_{0.2} \\
y & Y
\end{pmatrix}, \quad \text{and} \quad B = 
\begin{pmatrix}
b_{1.1} & b_{1.2} \\
 b_{2.1} & b_{2.2} \\
 \vdots & \vdots \\
b_{k.1} & b_{k.2}
\end{pmatrix},
$$

where $y_{0.1}$ and $y_{0.2}$ are intercepts. The least squares solution can be written as $Y = \Sigma^{-1}_a \Sigma_{ab}$ and $[y_{0.1} \ y_{0.2}] = \mu_b - \mu_a Y$ from the mean $\mu = [\mu_a \ \mu_b]$ and covariance matrix $\Sigma$ of the joint $(a, b)$ variables [3]. The mean and covariance matrix were replaced by MCD estimates ($\hat{\mu}$ and $\hat{\Sigma}$) [1] in order to yield the robust estimate of $\hat{Y}$. Note that this MCD estimates deal with noise in $A$ as well as $B$. The missing values in $g$ can be estimated by

$$
\begin{align*}
\alpha_1 & \approx y_{0.1} + w_1 y_{1.1} + w_2 y_{2.1} + \cdots + w_{n-2} y_{n-2.1}, \\
\alpha_2 & \approx y_{0.2} + w_1 y_{1.2} + w_2 y_{2.2} + \cdots + w_{n-2} y_{n-2.2}.
\end{align*}
$$

3 Experimental Results

We used gene expression ratio data from breast cancer experiments [4]. Gene expression ratios included in the data were derived from the fluorescent intensity (proportional to the gene expression level) from a tumor sample divided by the fluorescent intensity from a common reference sample (MCF-10A cell line) [4]. We prepared a BRCA2 mutation dataset (3,226 genes, 8 samples). There is no zero or negative value in this dataset. This is a challenging dataset since it is
not easy to impute missing values in cancer gene expression datasets built from various tumor tissues and/or different patients. The minimal value and maximal value of this dataset were 0.01 and 5530.7 respectively. The mean value and standard deviation were $\mu_{BRCA2} = 2.9$ and $\sigma_{BRCA2} = 62.3$ respectively. The BRCA2 dataset contains some big values including 5530.7 and 5362.2. One may be able to reduce the effect of outliers by log-transformation. However, we used the BRCA2 dataset without log-transformation to demonstrate outlier effect clearly.

We randomly deleted a total of $t$ entries of a given data matrix to make a missing matrix and $t$ answer values. The performance of the missing value estimation is evaluated by the mean relative error (MRE):

$$\text{MRE} = \frac{1}{t} \sum_{i=1}^{t} \frac{|\alpha_i - g_i|}{|g_i|},$$  \hspace{1cm} (2)

where $\alpha_i$ and $g_i$ are an estimated value and a known answer value, respectively, for the $i$-th missing entry ($1 \leq i \leq t$). After repeating three times randomly deleting the same percentage of entries of a given complete matrix, we can obtain three different sets of locations of missing values. Then, we report the average of MREs obtained from three simulations of imputation. We compared our method with KNNimpute [5], BPCA [6], and LLSimpute [2].

![Graph showing comparison of MREs for four methods](image)

**Fig. 1.** Comparison of the average MREs against 1% of missing entries for four methods (KNNimpute [5], BPCA [6], LLSimpute [2], and LLSimpute/MCD) on the BRCA2 dataset (3,226 genes, 8 samples). The average MREs for BPCA and LLSimpute with large $k$ were larger than 1.0 (not drawn).

In Figure 1, we compared the average MREs on the BRCA2 dataset (3,226 genes, 8 samples) with 1% of missing entries. The average MRE of BPCA was
about 1.8. The average MREs of LLSimpute with large $k$ values were also larger than 1.0. LLSimpute/MCD outperformed the other methods for the BRCA2 dataset that has outliers. When there is no obvious outlier in a dataset, there is no guarantee that LLSimpute/MCD can show better performance than LLSimpute. The optimal $k$ value for LLSimpute/MCD can be determined by the same procedure presented in [2].

4 Summary

We introduced an imputation method based on robust regression using MCD estimates. This method reduced outlier effect and showed better performance than other missing value estimation methods we tested for a cancer gene expression dataset.

Acknowledgments. This material is based upon work supported in part by the National Science Foundation Grants ACI-0305543 and CCF-0621889. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Design Pattern for Protein Identification

Jens Lichtenberg1, Lonnie Welch1,2,3

1 School of Electrical Engineering and Computer Science,
2 Biomedical Engineering Program,
3 Molecular and Cellular Biology,
Ohio University,
Athens, Ohio, 45701, USA
{lichtenj,welch}@ohio.edu

Abstract. This paper presents the protein identification problem in form of a design pattern [2]. Not unlike biological patterns, who describe reoccurring structures within biological objects, a design pattern in regard of software engineering is seen as the description of a thing, which is alive and the process, which creates it. Christopher Alexander [1] defines patterns as things that happen in the world and rules, which tell how to create such things and when they must be created. The pattern presented in this article addresses the identification of the correct protein based on a given peak list. It provides a process on how to infer the identity and primary structure of such protein based on searches of existing protein mass databases and a process describing how to create these databases. The current standards in proteomics describe data input and reporting between various analysis tool, mainly focused on protein identification and protein-protein interaction. The HUPO Proteomics Standards Initiative (PSI) [3] defines standards for data representation in proteomics to facilitate data comparison, exchange and verification. This goes along with common definitions of data formats in computer science as specific arrangements of data printed on a page, stored in a record, data file or storage device. A pattern however is defined as a plan, model or description to be followed in creating things and processes that are alive or generative.
The pattern presented in this paper is build on the understanding of the database search algorithms incorporated in various existing mass spectrometry based protein identification tools. It describes how protein identification algorithms compare experimentally derived mass data in form of peak lists to in-silico created mass information of proteins. It elaborates on the description of problems during the comparison of mass information and how to overcome them. The design pattern for protein identification enables software developers to create their own protein mass fingerprinting solutions, which can be adjusted, modified and extended to an experiment’s specific needs. To visualize and enhance the understanding of the pattern a process flowchart of protein identification is presented as well as a local implementation of the pattern.

Keywords: Pattern, Protein Identification, mzdata, Software Engineering, HUPO, PSI
References


Characterizing Pseudobase and predicting RNA secondary structure with simple H-type pseudoknots

Oyun-Erdene Namsrai, Kwang Jung, Sunshin Kim, Chung Rhee, Keun Ryu †

School of Electrical & Computer Engineering,
Chungbuk National University
Email: {oyunerdene, ksjung, sskim04, khryu}@dblab.chungbuk.ac.kr
csrhee@chungbuk.ac.kr

Abstract. After we had analyzed all the pseudoknots in Pseudobase, we classified it into six basic pseudoknot types. Of 245 unique pseudoknots in PseudoBase database, 235 were the H-type pseudoknots. Our characterization of Pseudobase shows that efficient and effective tools for detecting the RNA H-type pseudoknots are needed. In this paper, we present a computation the maximum number of base pairs of an RNA sequence with simple H-type pseudoknots. Our algorithm called “Iterated Base Pair Maximizing” is based on the Nussinov’s base pair maximizing algorithm. To evaluate our method we use the more than 20 sequences with simple H-type pseudoknots of variable size from 19 to 30 nucleotides. We get our experimental data set from PseudoBase, the algorithm correctly identifies more than 75% of base pairs for short sequences and 65% overall. It correctly predicts nearly all H-type pseudoknots. In addition, our approach can be easily extended and applied to other classes of more general pseudoknots. Availability: The algorithm has been implemented in C++ in a program called IBPM, which is available at http://dblab.cbu.ac.kr/ibpm

1. Characterizing Pseudobase and H-type Pseudoknots

PseudoBase is a database containing structural, functional and sequence data related to RNA pseudoknots [1].

As of May 2006, there are 245 unique pseudoknots in PseudoBase database. Of these, 189 or 77% are the simple “ABAB or ( [ ) ] ” H-type pseudoknots, 12 or 4.9% are the LL type pseudoknots (as shown Table1.), 24 or 9.8% are the HLout type, 11 or 4.5% are the HLin type, 2 are double pseudoknotted structure. Loop2 is often very short, in our observation 173 of the 189 unique simple H-type pseudoknots have Zero length, 187 of the 245 or 80% have Loop2≤1. After we had analyzed all the pseudoknots in Pseudobase, we classified it into six basic pseudoknot types as shown in Table1. and Fig. 1.

† Corresponding Author
* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (The Regional Research Universities Program/Chungbuk BIT Research–Oriented University Consortium)
Table 1. Classification of 245 pseudoknots in Pseudobase.

<table>
<thead>
<tr>
<th>PKB#</th>
<th>Name</th>
<th>Shortened Bracket View</th>
<th># of occ.</th>
<th>% of occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loop in</td>
<td></td>
<td>12</td>
<td>4.9</td>
</tr>
<tr>
<td>174, 140, 144, 145, 146, 143, 142, 141, 57, 139, 150, 76</td>
<td>RSV, CGMMV_PKbulge, ORSV-S1_PKbulge, RSV-S1_PKbulge3, Pseudovirus_PKbulge, STMV_PKbulge, TMGMV_PKbulge, TMV_PKbulge, TMV-L_Pkbolge, Ec_RNaseP-P6, HDV-It_ag</td>
<td>(((()))), (((())))</td>
<td>12</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>Hairpin &amp; Loop out</td>
<td></td>
<td>24</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>Hairpin &amp; Lin</td>
<td></td>
<td>11</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>HH</td>
<td></td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>181</td>
<td>HCV_IRE</td>
<td>((()))</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>HHH</td>
<td></td>
<td>6</td>
<td>2.4</td>
</tr>
<tr>
<td>171, 169, 178, 173*, 150, 163</td>
<td>HCV-229E, CoxB3, Ni_VS, satRPV*, Ec_RNaseP-P6, Hs_SR-Pkn</td>
<td>(((()))), (((()))), (((()))), (((()))), (((())))</td>
<td>6</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>All others (ABAB or ( ) )</td>
<td></td>
<td>189</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>245</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. Representations of Pseudoknot types: (A) ABAB type, (B) LL type, (C) HLout type, (D) HLin type, (E) HH type, (F) HHH type. H: hairpin, L: internal, Bulge or Multiple loop.

2 Our Approach

Our algorithm is based on the base pair maximizing algorithm [2][3], which we will describe briefly first. The Nussinov-Jacobsen algorithm, based on the following
recurrence, computes, for each $1 \leq i < j \leq n$ the quantity $B[i, j]$ which is the maximum number of pairs in any folding of the substring $x_i x_{i+1} \ldots x_j$ of the input:

$$B[i, j] = 0,$$ if $i \geq j - 4$

$$\max \{ B[i+1, j-1] + p_{i,j}, \max \{ B[i,k] + B[k+1, j] : i \leq k < j \} \} , \text{ if } i < j - 4,$$

where $p_{i,j} = 1$ if $x_i$ could pair with $x_j$ (i.e., G-C or A-U pairs),

and $p_{i,j} = 0$ otherwise (1)

Nussinov algorithm considers only regular base pairs. To compensate this problem and to get more good result when predicting secondary structures, we also considered G-U base pairing and we give different weights to $p_{i,j}$ defending on energy rules[4].

We now extend the basic base pair maximizing algorithm to accommodate H-type pseudoknots. Therefore, we can run the Base Pair Maximizing algorithm twice to identify H-type of pseudoknots. Firstly we predict a secondary structure of simple H-type pseudoknot using the Base Pair Maximizing algorithm. Secondly we remove all the paired bases and predict remaining sequence. H-type pseudoknots the only type of pseudoknot for which an energy model exists. Consequently, combining energy minimization method with IBPM method may give us good results.

A. $( [ ] ) = ( ) + [ ]$

B. $ABAB = AA + BB$

C. Fig. 2. A pseudoknot can be treated as two separate helices. First helices is identified by traditional BPM, second one is identified by IBPM algorithm.

**Our algorithm sketch as follows:**

1. Prepare input sequence. It should contain RNA sequence only.
2. Run the Base pair maximizing (or energy minimization method ) method and retrieve the base-pairing score matrix
3. Write to output base pair list, which are found from the score matrix.
4. Eliminate those base pairs and repeat steps 2-4 until no bases remain.
5. Merge output results and write it to final output.

Similarly, more complicated pseudoknots as shown in Fig. 1 (HHH, HH, HLin, Hlout type ) may identified with more iterations.

We run the IBPM algorithm multiple times, and each time we only accept the base-pairs that appear to be the most reliable, e.g. with the highest score. This modification attempts to avoid possible false predictions.

**6. Experimental Review and Discussion**

We predicted the structure of 12 pseudoknotted RNA sequences taken from PseudoBase. Average lengths of our experimental sequences taken from PseudoBase were around 20 nucleotides.
Table 2. Some output of our IBPM, PKB # is the Sequence ID on Pseu toxinBase database, PBase denotes the bracket view of structure on Pseu toxinBase.

<table>
<thead>
<tr>
<th>PKB #</th>
<th>PKB37</th>
<th>PKB39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>CCCCUUACUUGAGGGAAAUCAAGC</td>
<td>CCCCUUACUUGAGGGAAAUCAAGC</td>
</tr>
<tr>
<td>PBase</td>
<td>[: (:: :: [: [[ ]] :: :: ] ] ]</td>
<td>[: (:: :: [ :: :: ] ] ]</td>
</tr>
<tr>
<td>Prediction</td>
<td>[: (:: :: :: [ :: :: ] ] ]</td>
<td>[: (:: :: :: [ :: :: ] ] ]</td>
</tr>
</tbody>
</table>

To evaluate our software we compared our result to that of the current best known software for globally optimal RNA pseudoknot prediction developed by Elena Rivas. That algorithm has a worst case complexity of O(N^6) in time and O(N^4) in storage. Our approach based on traditional Base Pair Maximizing technique proposed by Nussinov [1]. The algorithm has worst case time and space complexities of O(N^3) and O(N^2), respectively.

IBPM folds 12 pseudoknots out of 12 in the Pseudobase database. Our program predicts simple pseudoknots with correct or almost correct structure for 65% sequences. Table 2. Shows some output of our IBPM program. From 120 pairs which are included in our total 12 pseudoknotted sequences, we predicted total 78 base pairs.

In this paper, we analyzed whole pseudoknots stored in Pseudobase Database. We classified it into sex main categories and represented graphical representations of them. Therefore we observed that H-type pseudoknots are the most abundant of all known pseudoknots. We also presented an algorithm for RNA secondary structure prediction with H-type pseudoknots, based on the traditional base pair maximizing algorithm. But due to the space complexity of dynamic programming technique for maximizing the number of base pairs, our algorithm cannot deal with longer sequences.

References

Mining MEDLINE for Gene Clustering: A Comparison of Feature Selection Approaches

Sailaja Pydimarri, Orlando Karam, Venu Dasigi, and Rajnish Singh
Southern Polytechnic State University (Marietta, GA)

Abstract. In this paper we compare two different schemas for feature selection applied to data mining of MEDLINE data. We also highlight the importance of the choice of background set for both of these approaches. We compare the performance of these schemas and the choice of background sets both by keyword evaluation by an expert and by applying it to document clustering.

1 Introduction

There is extensive work on algorithms for mining applications, many of them applied to biomedical literature. Statistically, keywords have been defined as those words that occur with a substantially larger frequency in the documents specific to a protein family compared to a general background set of documents. This is captured by a statistical measure called z-score, and words with large z-scores are important and distinctive of the concept in consideration [1].

An alternative metric, called TF-IDF, uses the product of normalized frequency of a term (TF) in a document and the inverse document frequency of the term (IDF) to weight a term in a document. The idea is that terms that capture the content of a document occur frequently in the document (that is, their TF is high), but if such a term were to be a good term that discriminates the document from others, it occurs in only a few documents in the general population (that is, its IDF should be high, as well).

Liu, et al. evaluated the TF-IDF and z-score schemes in identifying keywords. They used the keywords thus identified in their gene clustering experiments, where clusters were created based on shared keywords. Keywords based on the TF-IDF metric resulted in higher quality clusters than those based on the z-score [1].

In this paper we perform experiments similar to those reported in [1], with a few important differences. Our ultimate goal is to not just compare, but combine the TF-IDF and z-score approaches. Toward this end, we initially tried to replicate their experiments, on the version of MEDLINE available to us. As we do this, we strive to fully specify our choice of parameters, to make it easier for others to reproduce our experiments. We also explore the impact of several parameters they didn’t explore, including the choice of background set and the choice of different weighting schemes.
2 The Problem and the Methodology

In this work, we approach the gene clustering problem as follows. We start with a set of 44 yeast genes, described in [2], and seek to cluster them into groups of functionally related genes. We compare our clustering against the standard cited in [2]. The features used by the clustering algorithm are keywords selected based on their properties, such as either the TF-IDF or the z-score metrics. All articles from MEDLINE that mention each of the genes in question are considered to form a group of documents related to the gene. When we refer to a group of documents below, they relate to a specific gene.

3 Feature Selection approaches

For selecting which particular features (i.e. words) are important for each particular group of documents, we tested two different approaches, TF/IDF and z-scores, modified to apply to documents within a group rather than directly to words within a document. The following sections describe the particular formulas we used.

\[ TF_g^a = \frac{df_g^a}{|g|} \]  \hspace{1cm} (1)

\[ idf^a = \log \frac{N}{df^a} \]  \hspace{1cm} (2)

where \( df_g^a \) refers to the number of groups in the background set in which word \( a \) occurs, and \( N \) is the total number of groups in the background set.

The z-score of a word \( a \) for a group \( g \) is defined as:

\[ Z_g^a = \frac{F_g^a - \overline{F^a}}{\sigma^a} \]  \hspace{1cm} (3)

where \( F_g^a \) equals the proportion of documents in group \( g \) that contain the word \( a \), \( \overline{F^a} \) is the mean of \( F_g^a \) across all groups and \( \sigma^a \) the standard deviation of \( F_g^a \) across all groups.

Choosing the background set

Another important parameter we tested was the choice of the background set. Both attribute selection schemes intuitively rely on comparing the prevalence (frequency) of a term within a certain group of documents, as compared to its prevalence in another set, which we call the background set. For TF/IDF the Document Frequency (DF) and consequently its inverse, the IDF are obtained from the background set (traditionally the full set of documents available). For the z-score, the background set determines both the mean frequency and the standard deviation.

Intuitively, two possible background sets appear interesting:
– **The full MEDLINE set** This is the most commonly used set for TF-IDF. It requires the documents to be divided into random groups for calculating the z-score. Intuitively, we are comparing the prevalence of the word in the interest group as compared to the word’s prevalence within the whole universe of documents.

– **The union of all groups** This is the set of all documents that belong in one or more of the groups of interest. For this set, the groups for z-score are the naturally formed groups. Intuitively, we are comparing the prevalence of the word in a group as compared to its prevalence in all the groups. Words that appear in all the groups with a higher frequency than in the full MEDLINE set, but which are not more prevalent in a particular group will get lower scores than with the full MEDLINE set.

We hypothesize that keywords obtained using the full MEDLINE set are better for clustering, while those obtained using the union of groups more accurately characterize the group (in our case, the gene).

## 4 Experimental Comparison

### 4.1 Expert evaluation of keyword quality

We used all four schemes to generate a list of the top 30 words for each schema for 3 of the genes (ace2, cdc21 and mnn1). According to the expert, words obtained using the TF-IDF produced better keywords than z-score. The background set didn’t appear to greatly affect the quality in any case. An interesting observation was made for ace2, a word that refers to two different genes, a yeast gene and a human one. When using the union of groups as the background for z-score, more keywords related to the function of the human gene (renal activity) were selected than with the full MEDLINE set.

### 4.2 Clustering

We performed k-means clustering on the groups using the features obtained by calculating the union of the top scoring words for each group. We then applied k-means clustering to the resulting data set, applying a weighting scheme. We evaluated the clusters by calculating their purity as compared with an expert-generated clustering for those genes.

We tried several schemes for assigning weights to the features for each group. The features are the words. The weights are functions of some properties of those features, where the properties may be things like term frequency, document frequency, et cetera. We tried using the raw z-scores and TF-IDF scores as the weight for the feature, with disappointing results. We then tried converting the property to binary values, for both frequency (as a surrogate of TF-IDF) and z-scores, with much better results. The best results were obtained when converting frequency to binary values, assigning a 1 if the word appears in at
least one document within the group and 0 otherwise. Notice this is equivalent, but simpler, to assigning 1 or 0 depending on the TF-IDF being 0 or not.

We used the binarized frequency as the weight for our feature for all the experiments we report.

For the k-means clustering we used 9 clusters (the same number of clusters as those present in the expert clustering), and ran the algorithm with 1000 iterations. We used the Cluster 3.0 software, described in [3].

Given that for clustering, words that appear in only one group would not add much information (other than to separate the group from everybody else), we eliminate all words that do not appear in at least two groups.

Fig 1 shows the purity of the clusters obtained by running the k-means algorithm using only the features that are the top n for at least one group.

<table>
<thead>
<tr>
<th>Algorithm—Background</th>
<th>Top 10—#words</th>
<th>Top 30—#words</th>
<th>Top 50—#Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF — union of groups</td>
<td>.636—315</td>
<td>.681—830</td>
<td>.561—1383</td>
</tr>
<tr>
<td>TF-IDF — full MEDLINE</td>
<td>.681—247</td>
<td>.681—590</td>
<td>.659—885</td>
</tr>
<tr>
<td>z-score — union of groups</td>
<td>.409—475</td>
<td>.477—1524</td>
<td>.431—2280</td>
</tr>
<tr>
<td>z-score — full MEDLINE</td>
<td>.681—309</td>
<td>.613—747</td>
<td>.636—1139</td>
</tr>
</tbody>
</table>

Fig. 1. Purity of clusters by algorithm used and cutoff rank

From this table we can notice that the z-score using the foreground set as the background set also provides the worst performance. We found that TF-IDF produced good results with either background set. Differently to [1] we found z-score performance to be comparable to TF-IDF when the background set is the full set of documents.

If we also take into consideration the number of features selected, TF-IDF with the full MEDLINE set has a small advantage, using 247 words against 309 for z-score with the same background set. It appears that the full MEDLINE set should be used as the background for purposes of clustering.

References

Evaluation of Stability Changes in Single Point Protein Mutants Using a Four-body Statistical Potential

Gregory M. Reck and Iosif I. Vaisman
College of Science, George Mason University
10900 University Blvd., Manassas, VA 20109
greck@gmu.edu, ivaisman@gmu.edu

Abstract. A statistical potential function has been derived based on Delaunay tessellation of a large number of computationally hydrated protein structures for eventual use in mutant stability studies. The tessellation provides quadruplets of “nearest neighbor” residues and the potential function is derived from the observed occurrence of each quadruplet compared to a reference expectation. Several potential functions are computed from the hydrated structures using alternative reference expectations and compared with a function from unhydrated structures using a random distribution expectation. Each of the potential functions is used to score groups of single point mutants of barnase, staphylococcal nuclease and T4 lysozyme, and the scores are correlated with reported stability measurements of the mutant proteins. Three machine learning tools (decision tree, random forests, support vector machine) are used to demonstrate a relationship between the potential function scores and the stabilities of the mutants.

1 Introduction

Knowledge-based models incorporating statistical potentials derived from atom or residue contacts in protein structure files have been used for structure analysis and folding prediction [1]. Studies have often relied on the Boltzmann principle to link contact frequencies to mean force potentials [2]. Many efforts have used pairwise contact frequencies, but this study is based on 4-body relationships defined by Delaunay tessellation of the residue locations [3]. This computational geometry method reduces the convex volume surrounding the residues into a set of space-filling irregular tetrahedra. Each tetrahedron has a residue location at its 4 vertices and these residues are defined as 4-body nearest neighbors. The statistical potential is derived from analysis of a large set of protein structures and expresses the log likelihood of occurrence of each possible quadruplet compared to a reference expectation.

The protein environment plays a critical role in guiding protein folding, and previous contact potential studies have used various techniques to represent the water milieu typical of globular proteins. This study explicitly computes plausible water positions in a shell surrounding the protein, and these waters are incorporated into the tessellation. Since water molecules are smaller and not randomly distributed with respect to the residues, two alternatives are examined: a split potential (different poten-
tials for core and surface residues), and a water group potential (the expected occurrence is based on the water content of the quadruplets).

When these statistical potentials are applied to a new protein, they provide a total potential (sum of individual quadruplet potentials) that expresses the extent of compatibility between the protein sequence and the protein structure [4]. This hypothesis is tested by computing the total potential for three proteins with large experimental stability databases of single point mutations. The changes in potential are correlated with changes in stability associated with each of the mutants. Correlations are attempted with conserved, nonconserved and other specific residue types. Summing the potentials of the tetrahedra surrounding each residue can provide a potential profile along the protein main chain. Changes in potential profiles for mutated proteins (residual profiles) are used to train three machine learning tools, and the resulting classifiers are evaluated using standard repetitive, 10-fold cross validation techniques.

2 Materials & Methods

A reference set of 1417 nonhomologous, high resolution protein structure files were selected from the PDB for derivations of the unhydrated potentials, and 1353 of these were eventually used to develop the hydrated protein potentials. A full complement of hydrogen atoms were added to the structure files using MolProbity [5] before hydration and then SOLVATE [6] was used to locate plausible water positions around each protein with a minimum layer thickness of 5 Å. The protein structures were tessellated using the QuickHull program [7] based on the carbon alpha positions of the residues. For hydrated protein files, all SOLVATE water placements were included in the tessellation, but any tetrahedra comprised of only water locations were then discarded. The resulting sets of tetrahedra (or residue quadruplets) were used to develop the statistical potentials.

The statistical potential functions were extracted from the assembly of quadruplets from all of the reference proteins. Equation 1 shows the derivation of the unhydrated potential (CA) for each possible \( i, j, k, l \) quadruplet.

\[
q_{ijkl} = \log \left( \frac{f_{ijkl}}{p_{ijkl}} \right) = \log \left( \frac{f_{ijkl}}{ca_i a_j a_k a_l} \right).
\]

(1)

where \( f_{ijkl} \) is the observed frequency of occurrence for the quadruplet, \( p_{ijkl} \) is the expected frequency of occurrence (both normalized), \( a \) is the frequency of the specific residue, and \( c \) is the permutation factor shown in equation 2.

\[
c = 4! \prod_{i=1}^{r} \left( t_i! \right),
\]

(2)

where \( t \) is the number of times that a residue of type \( i \) appears in the quadruplet and \( r \) is the total number of residue types in the reference set of proteins.

For the reference set of hydrated proteins, the split potential (SP) is created by separating the quadruplets into a core set and a surface set, and using the equations above to derive a potential for each. A residue is defined as core when more than half
of the surrounding tetrahedra have 0 or 1 water molecules, and surface when more than half have 2 or 3 waters. The water group potential is shown in equation (3)

\[ q_{ijkl} = \log \left( f_{ijkl} / p_m [c_n a_n] \right) \]  

(3)

where \( m \) is the number of waters in the quadruplet, \( p_m \) is the probability that a quadruplet has \( m \) waters, and the permutation factor and residue probability are applied only to the \( n \) natural residues in the quadruplet.

Total potential scores for mutant proteins are determined by changing the identity of the mutant residue and recomputing the potentials. This assumes that the mutation does not significantly alter the structure. The ProTherm [8] database facility was used to identify experimental single point mutant stability change data for 132 mutants of barnase, 497 of staphylococcal nuclease and 377 of T4 lysozyme over 4 pH levels.

The WEKA environment [9] was used for the three machine learning tools: decision tree (DT), random forest (RF) and support vector machine (SVM). Each individual mutant residual profile served as an instance and the stability values were discretized. The classifiers were evaluated using percent success, the Kappa statistic, and the Area Under the Receiver Operating Curve (AUC).

3 Results

Each of the experimental datasets was correlated with the corresponding mutant residual potentials, and the results are shown in Table 1. The correlation coefficients for barnase and nuclease are smaller than for lysozyme that ranges up to 0.7. All are significant except for barnase with the WG potential. Overall, the SP potential outperforms the other potentials and WG has the lowest correlation values. The \( r \)-values for nonconserved mutants are frequently larger than conserved for all 3 proteins.

<table>
<thead>
<tr>
<th>Protein</th>
<th>pH</th>
<th>Statistic</th>
<th>CA</th>
<th>WG</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnase</td>
<td>6.3</td>
<td>( r )</td>
<td>0.2803</td>
<td>0.0418</td>
<td>0.1798</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td>0.0011</td>
<td>0.6338</td>
<td>0.0391</td>
</tr>
<tr>
<td>S. nuclease</td>
<td>7.0</td>
<td>( r )</td>
<td><strong>0.2806</strong></td>
<td><strong>0.2160</strong></td>
<td><strong>0.4559</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td>1.9E-12</td>
<td>7.7E-08</td>
<td>&lt;2.2E-16</td>
</tr>
<tr>
<td>T4 lysozyme</td>
<td>2.0</td>
<td>( r )</td>
<td><strong>0.3750</strong></td>
<td><strong>0.3044</strong></td>
<td><strong>0.4180</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td>8.4E-04</td>
<td>0.0075</td>
<td>1.7E-04</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>( r )</td>
<td><strong>0.4582</strong></td>
<td><strong>0.3972</strong></td>
<td><strong>0.4931</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td>3.7E-08</td>
<td>2.6E-06</td>
<td>2.1E-09</td>
</tr>
<tr>
<td></td>
<td>5.5</td>
<td>( r )</td>
<td><strong>0.6697</strong></td>
<td><strong>0.6598</strong></td>
<td><strong>0.7144</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td>3.4E-14</td>
<td>1.1E-13</td>
<td>&lt;2.2E-16</td>
</tr>
<tr>
<td></td>
<td>6.5</td>
<td>( r )</td>
<td><strong>0.5751</strong></td>
<td><strong>0.4955</strong></td>
<td><strong>0.6044</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td>4.4E-07</td>
<td>2.3E-05</td>
<td>7.6E-08</td>
</tr>
</tbody>
</table>

Table 1. Pearson correlation coefficients (\( r \)) and p-values for single point mutants of 3 proteins. Values shown in bold are statistically significant (\( p < 0.05 \)).
Several specific groups of nuclease mutants were also examined. The SP potential achieved highly significant r values between 0.6 and 0.7 for a group of 82 mutants of large hydrophobic residues, 56 mutants that make or break H-bonds, and a residue scan at a beta-turn position near the surface. The r values for a group of 74 mutants of polar uncharged residues and 22 mutants that reverse or neutralize a surface charge were slightly lower but highly significant. Correlations with 95 mutants of ionizable residues to alanine and glycine, and 50 surface residues to phenylalanine were not significant.

Residual profiles (mutant – wild type) nuclease data were used to train the ML classifiers and the results were quite positive. The RF program performed best with success rates over 75%, Kappa statistics over 0.5, and AUC values over 0.8. These values are all well above the random classifier level. Results were higher when attributes such as mutant identity and location were added. Results were lower, but still positive when the extent of discretization of the stability attribute was increased.

4 Conclusions

The results indicate that statistical potentials derived from Delaunay tessellation of computationally hydrated protein structure files can be successfully correlated with experimental data on the stability of single point mutations. This indicates that the potentials identify sequence-structure compatibility and can be effective in studies of protein structure and folding.

References

A Topological Characterization of Protein-Water Interactions for Knowledge-Based Models

Gregory M. Reck and Iosif I. Vaisman
College of Science, George Mason University
10900 University Blvd., Manassas, VA 20109
greck@gmu.edu, ivaisman@gmu.edu

Abstract. A new strategy is described that characterizes the relationship between a globular protein and its water environment. This technique uses the 4-body nearest-neighbor relationships defined by Delaunay tessellation of computationally hydrated proteins. The new water group parameter is based on the extent of inclusion of simulated water molecules into the Delaunay tetrahedra surrounding each residue. Topological parameters are especially informative in developing knowledge-based methods for predicting protein structure and correlating protein behavior such as stability and functionality.

1 Introduction

Knowledge-based models have been suggested as an alternative to more computationally intensive physics-based simulations for the study of protein stability and folding [1–4]. These methods typically employ empirical information derived from high-resolution, experimentally determined protein structure databases such as contact frequencies or spatial relationships. Choosing informative parameters that characterize the relevant interactions in protein systems is essential.

Since the solvent surrounding a nascent protein is a crucial element in the folding process, parameters that capture characteristics of protein residue interactions with water are essential in studies of protein structure. These include features such as residue depth with respect to the surface, and the extent of contact or accessibility to water. These parameters may also be useful in classifying residues into subsets for component characterizations. Solvent accessible surface area [5] is frequently used for this purpose, but other parameters such as circular variance [6], residue depth [7], and water coordination number [8] may be more effective for partially buried residues.

This study examines two strategies for surface characterization, both based on Delaunay tessellation (DT). This computational geometry technique reduces the space around the residue coordinates of a protein into a set of irregular tetrahedra (or simplices) where each tetrahedron is defined by a residue location at its four vertices. Since there are no other residues within the tetrahedron, the four residues can be unambiguously defined as nearest neighbors without resort to arbitrary criteria such as cutoff distances. The first classification scheme uses the PDB structure files directly (unhydrated) and searches the simplices for unmatched faces that contain surface
residues [9]. The second is a new strategy that relies on computationally hydrated structures, and that provides a surface classification as well as a quantitative measure of solvent water influence on each residue. Both approaches are compared with other surface characterization parameters for a large set of protein structures.

2 Materials & Methods

A reference set of 1321 high resolution, non-homologous protein structure files was selected from the PDB. Since this study focuses on protein-water interactions, proteins classified as membrane proteins by DSSP [10] were not included. The SOLVATE program [11] was used to hydrate the PDB files. SOLVATE progressively places waters starting at the surface, with a limited optimization using van der Waals forces, but no electrostatic forces are included. Water is placed in a shell around the protein with a minimum thickness of 5 Å to capture solute effects in the first hydration shell [12]. The resulting water positions are not as realistic as other methods (molecular dynamics), but positions of oxygen atoms should be sufficiently accurate for the current tessellation-based study of the extent of water association with surface residues. DOWSER [13], another water placement program based on energy considerations was used to investigate the SOLVATE water placements in internal locations and external crevices for several proteins.

The protein structures were tessellated using the QuickHull program [14] based on the carbon alpha positions of the residues. For hydrated protein files, all SOLVATE water placements were included in the tessellation, but any tetrahedra comprised of only water locations were discarded. So only water molecules in the tessellation hydration layer around the protein remain, i.e., those that are linked to a residue by the edge of a tetrahedron.

The simplex face match (SFM) technique for surface residue classification uses a simple search to identify tetrahedra with unmatched (surface) faces. The three residues at the vertices of the triangular face are classified as surface. Residues linked to a surface residue, but not in a surface face are classified as under-surface, and the remaining residues are classified as buried.

The water group number (WGN) technique identifies and examines all the tetrahedra surrounding a residue and counts the number of tetrahedra with 0 water, 1 water, 2 waters and 3 waters. Then the water group number is computed as the least squares slope of the tetrahedra count vs. the water content. If the WGN is greater than 0.5 the residue is classified as surface, otherwise core.

Accessible surface area (ASA) was computed using the NACCESS program [15]. The water coordination number (WCN) is a variation of the atom coordination number [8] and is simply the number of nearest neighbor water molecules to a residue. Residue depth was computed using the DPX program [7] that finds the distance to the nearest solvent accessible residue. Finally, the SIMULAID program [6] was used to calculate the circular variance (CV) for each residue. CV is the angular spread of a group of vectors that are drawn from a residue query point to all other residues in the protein. A point with CV of zero is at the center of the set of points and CV of 0.5 indicates a surface point.
3 Results

Protein hydration has an important impact on the results of tessellation. When an unhydrated protein is tessellated, some simplices (tetrahedra) in regions of surface irregularities may be highly distorted. But when the same protein is hydrated and then tessellated, the surface waters interrupt the elongated simplices. For the set of 1321 unhydrated proteins, the maximum simplex edge length was 113 Å with a mean of 8.2 Å. After hydration, the maximum was reduced to 12.9 Å with a mean of 4.85 Å.

The SOLVATE water placements around the 1321 proteins were examined in detail and compared with data from referenced studies on hydration layer characteristics. A mean value of 14.3 square Å per HOH is consistent with several previously reported values, although a range of values has been cited. A comparison with a hydration study of citrate synthase [16] indicates that tessellation hydration is near the maximum hydration loading.

A detailed comparison was made of the crystal HOH identified in the PDB files with the SOLVATE waters. The average separation distance was 1.215 Å and less than 0.66% of the separations were larger than 2.5Å. While the large separations are a small segment, they are of interest since they likely indicate deeply buried crystal HOH without SOLVATE counterparts. DOWSER was applied to a small number of proteins with these larger separations and the results indicate that for internal cavities, DOWSER predictions agree well with crystal HOH locations, but not SOLVATE. For external crevices, SOLVATE water placements are closer. Finally, an examination of the water coordination number for all the residues indicates that very few residues have no association with a water molecule. There is sufficient convolution and internal cavities that even deeply buried residues manage a tessellation link to water.

The two classification methods were applied to all the residues (293,421) in the reference set and the methods were compared based on their classifications. The SFM method classifies more residues as core (190,271) than WGN (140,553). Even if the SFM undersurface and surface classes are combined, WGN still identifies far more surface residues (152,868) than SFM (103,150). Again, it is apparent that the WGN finds extensive water association for many of the residues classified as undersurface and some classified as buried by the SFM.

Next, each of the two classification methods was compared to the other topological parameters, using both histograms and boxplots. Both methods evidence distinctive differences between surface and core groups with all the parameters. The WGM in particular shows quite different ASA and CV profiles for each group and almost a complete separation for WCN. This suggests a relationship between WGM and WCN.

Each of the classification methods and parameters was applied to an example protein, human transthyretin (1bmz, chainA). Inverse relationships are clearly evident between ASA and residue depth, and between CV and WCN. The structure in the protein is also evident in the results of the SFM and WGN classification methods.

Since the WGN is an indication of water association with the residues, a boxplot was drawn of WGN values for each residue type. The mean values for each residue suggest a relationship with residue hydrophobicity. This was tested by correlating the WGN means with 17 reported hydrophobicity scales. The $R^2$ correlation coefficients ranged from 0.4 to 0.9, with an average of 0.67.
4 Conclusions

A new strategy based on Delaunay tessellation for describing the protein–water interactions around a computationally hydrated protein has been presented and compared with other topological parameters. The water group number has shown distinctive relationships with ASA, circular variance, residue depth and coordination number. Mean values of WGN computed for each residue in the reference set have correlated well with standard hydrophobicity scales. This indicates that the water group number can be useful in future applications of knowledge-based models.

References

Integrated Statistical and Association Rule Analyses of Time-Dependent Gene Co-expression Patterns

Sandra L. Rodriguez-Zas1,2, Younghee Ko3, and Bruce R. Southey3,4

1 Department of Animal Sciences, 2 Department of Statistics, 3 Department of Computer Science, 4 Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

{Sandra Rodriguez-Zas, rodrgzzs@uiuc.edu}

Abstract. An integrated approach that combines powerful statistical and data mining tools with functional annotation information was used to fully characterize gene expression data. First, filtering, normalization and linear modeling of gene expression observations are used to minimize the impact of technical or nuisance variation on gene expression intensities and identify genes with differential expression across conditions. Then, condition-dependent thresholds are applied to the adjusted gene expression patterns and quantitative association rules are obtained. Interesting relationships among genes are identified based on the simultaneous consideration of complementary quality rule criteria. The integrated approach exploited the advantages of hierarchical models and association rule, assisting in the interpretation of gene expression data and providing novel insights into several sociogenomic pathways.

Keywords: Mixed effects model, Association rules, Microarrays.

1 Introduction

The vast information provided by microarray gene expression techniques challenges the discovery of biologically meaningful and interesting relationships between genes. Association rules are a data mining technique that identifies interesting relationships between items [1]. When applied to microarray data, association rules are described with left- and right-hand side gene(s) that have different expression patterns across conditions and the expression of a gene at any one condition is considered a transaction [2]. An advantage of association rules over other techniques to identify co-regulated genes (i.e. clustering), is that a gene can be present in multiple rules.

Association rules is a powerful tool to uncover relationships between genes however, two major challenges have hindered the widespread use of this technique. First, failure to remove technical and biological variation in the microarray data can lead to biased association rules. Second, the effective implementation of association rules must address the potentially heavy computational burden of identifying all possible relationships among thousands of genes. The objective of this study was to develop a novel approach to identify relationships between genes based on gene
expression data using association rules while addressing the challenges of this methodology. This approach integrates statistical analysis, condition-dependent quantitative association rules, and functional information. We applied the novel approach to characterize complex gene pathways involved in maturation and social behavior using microarray gene expression data from honey bees.

2 Materials and Methods

Gene expression levels at multiple behavioral maturation ages in honey bee (*Apis mellifera*), a well established model organism in sociogenomic studies, were measured using microarrays [3]. Nurse honeybees were sampled at days 0, 4, 8, 12, 17 after adult emergence and forager honey bees were sampled at day 17 after adult emergence [4]. Measurements were obtained from 20 microarrays in a loop design that maximized the direct comparisons between consecutive maturation stages. Gene expression was measured in individual honey bee brains using a double-spotted 9K cDNA (element) microarray following standard protocols [4]. Control and low quality features were filtered from the data and the average of the duplicated spots within elements and microarray were further analyzed. The log2 intensity values were normalized using a linear logarithmic transformation [5] and centered.

The first component of our novel approach consisted in removing technical and biological noise. To accomplish this, the normalized gene expression data was analyzed using the two-stage model of Wolfinger et al. [6]. In the first stage the global effects of dye and microarray were removed from the transformed gene expression intensities. In the second stage maturation stage and dye were fixed effects and array was a random effect in an element-specific model. The statistical significance *P*-value was used to identify the elements with differentially expression across stages. The first component of our approach provided unique estimable functions, least-squares means (also known as adjusted means) of gene expression at the six maturation stages for each element that were then used to identify interesting relationships among genes.

In the second component of our approach, the formulation of the quantitative association rules was based on the *a priori* algorithm implemented by Creighton and Hanash [2] and each maturation stage was considered a transaction. The implementation of traditional association rules uses expressed or not-expressed observations, however the least square means are continuous. Furthermore, genes can exhibit a wide range of variation in expression and one single threshold would not be adequate to transform the least square mean expression of all the genes into Boolean variables. To address this heteroskedastic scenario, the least-squares means were standardized within element to a mean of zero and a variance of one. Standardized least-squares means exceeding the average across all elements within stage were assigned an “expressed” status, otherwise they were “not-expressed” status.

To address the computational demand challenge of the association rule component of our approach, the elements that share the same pattern of expression were then grouped together into item-sets or groups. The item-set of elements with expression levels that did not surpass the threshold at any one stage was discarded because this
item-set contains no information about the pattern of expression across maturation. The item-set where the expression levels surpassed the threshold at all stages was also discarded because it represents the superset of all expression patterns. Association rules were computed between all possible pairs of item-sets and each rule was evaluated by four complementary quality rule statistics, support, joint support, confidence and conviction [1] [2].

3 Results and Discussion

A total of 6848 elements remained in the data set after filtering of control and unreliable features. A total of 1752 elements had significant ($P$-value $< 10^{-4}$) differential expression across stages. Considering all elements (with and without significant differential expression across stages), there were 61 item-sets. The average number of elements per group was 84 and ranged from 1 to 426. There were 51 item-sets including only elements with significant differential expression and the median number of elements per item-set was 20. Most item-sets lacked elements assigned to official genes and hence the low median number of elements with Gene Ontology classifications.

Table 1 presents the distribution of the number of association rules for all observed combinations of confidence levels for the left and right hand side item-sets and expression status. The confidence of the association rules ranged from zero to one. The rules with zero confidence correspond to the rules that have zero support. The conviction of association rules ranged from zero to infinity where zero corresponded to the rules with zero support and infinity corresponds to the rules with a confidence of one (Table 1). The highest number of rules was observed when all relationships had a conviction of one, indicating that these items are expected to be independent of each other.

Interesting rules can be identified by the simultaneous consideration of complementary rule quality indicators. Rules with a confidence of one are expected to occur all the time in the dataset. For example, given the expression status at any one stage on the item-set in the right hand side, the item-set in the left hand side will have the same expression status. Consideration of the ratio of the confidences between two item sets helps identify rules in which one direction is more likely to occur than the other direction. Rules can be further restricted by requiring high support and high conviction. For example, there were 12 rules that had a high support and high confidence, however, only 10 rules also had a high conviction. Another 52 rules had high confidence, high conviction and high ratio but only a moderate joint support.

The computational cost involved in computing quantitative association rules in gene expression studies was reduced by only considering significantly expressed elements, pairwise relationships between elements, and binning elements that have the same expression pattern across stages. The effect of each step is as follows, there were 6848 elements that were analyzed which results in 46,888,256 association rules. Only 1752 elements were significantly expressed over time which results in 3,067,752 possible rules. The binning of elements result in a maximum of 62 item-sets (generally $2^{n}$ (number of conditions)) and 3,782 rules.
Table 1. Number of rules by support, confidence, conviction and ratio levels where a ratio of -
2 indicates zero support and a ratio of -1 indicates a confidence of one.

<table>
<thead>
<tr>
<th>Support</th>
<th>0.00</th>
<th>0.30</th>
<th>0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Rules</td>
<td>312</td>
<td>1852</td>
<td>286</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Confidence</th>
<th>0.00</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rules</td>
<td>312</td>
<td>578</td>
<td>719</td>
<td>576</td>
<td>265</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conviction</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>1.25</th>
<th>1.5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Rules</td>
<td>34</td>
<td>164</td>
<td>650</td>
<td>748</td>
<td>285</td>
<td>569</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ratio</th>
<th>-2.00</th>
<th>-1.00</th>
<th>1.00</th>
<th>1.25</th>
<th>1.50</th>
<th>2.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Rules</td>
<td>312</td>
<td>265</td>
<td>608</td>
<td>510</td>
<td>424</td>
<td>331</td>
</tr>
</tbody>
</table>

The novel combination of statistical analysis and data mining association rules
exploited the advantages of both approaches to assist the interpretation of microarray
gene expression data. The expression of elements was filtered, adjusted for nuisance
terms including dye and microarray and the elements statistically differentially
expressed across conditions were identified. Subsequently, quantitative association
rules were applied to standardized least square means estimates at each stage that
characterize relationships between groups of elements with distinct pattern of
expression were identified.

Acknowledgments. Support of NIH/NIGMS grant 5R01GM068946 and NIH/NIDA
grant 5P30DA018310-039003 are highly appreciated.

References

1. Dunham, M.H.: Data mining. Introductory and advanced topics. Prentice Hall,
2. Creighton, C., Hanash, S.: Mining gene expression databases for association
3. Robinson, G.E.: Genomics and integrative analyses of division of labor in honey
4. Whitfield, C.W., Ben-Shahar, Y., Brillet, C., Leoncini, I., Crauser, D., Leconte,
   16075
6. Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H.,
   Bushel, P., Afshari, C., Paules, R.S.: Assessing gene significance from cDNA
   microarray expression data via mixed models. J. Comput. Biol. 8 (2001) 625-
   637
A Bioinformatics Approach to the Identification, Classification, and Analysis of Plant Hydroxyproline-Rich Glycoproteins

Allan M. Showalter\textsuperscript{1,2}, Brian Keppler\textsuperscript{1}, Jason Yerardi\textsuperscript{3}, Tom Conley\textsuperscript{3}, Lonnie R. Welch\textsuperscript{2,4,}, and Dazhang Gu\textsuperscript{3}

\textsuperscript{1} Department of Environmental and Plant Biology
\textsuperscript{2} Molecular and Cellular Biology Program
\textsuperscript{3} School of Electrical Engineering and Computer Science
\textsuperscript{4} Biomedical Engineering Program
Ohio University, Athens, OH 45701

Hydroxyproline-rich glycoproteins (HRGPs) are a superfamily of plant cell wall proteins implicated to function in diverse aspects of plant growth and development. The HRGP superfamily consists of three members: the hyperglycosylated arabinogalactan-proteins (AGPs), the moderately glycosylated extensins (EXTs), and the lightly glycosylated proline-rich proteins (PRPs). In addition, hybrid and chimeric versions of these HRGPs also exist within the HRGP superfamily.

In order to "mine" genomic databases for HRGP family members and to facilitate and guide research in the field, a "BIO OHIO" software program was developed that identifies and classifies AGPs, EXTs, PRPs, hybrid HRGPs, and chimeric HRGPs from proteins predicted from DNA sequence data. This bioinformatics program is based in part on searching for biased amino acid compositions (e.g., 50\% of an AGP’s amino acid composition consists of proline (P), alanine (A), serine (S), and threonine (T)) and in part on searching for particular protein motifs associated with known HRGPs (e.g., the pentapeptide sequence SP4 and tetrapeptide sequence SP3 are characteristic of EXTs). In addition, potential signal peptide sequences (which are generally associated with HRGPs and the secretory pathway) and glycosylphosphatidylinositol (GPI) lipid anchors (which are frequently associated with AGPs and responsible for their transient plasma membrane localization) are identified by the program. HRGPs identified by the program are subsequently analyzed to: 1) reveal any novel repeating amino acid sequences, 2) create protein phylogenies, 3) elucidate expression patterns of their genes using public databases, 4) determine whether genetic mutants exist for their genes, and 5) elucidate putative glycosyl transferase, prolyl hydroxylase, and peroxidase genes coexpressed with their genes in Arabidopsis.

To date, the BIO OHIO software was used to identify and classify 140 HRGPs from the Arabidopsis proteome as follows: 65 AGPs (including 20 classical AGPs, 3 lysine-rich AGPs, 15 arabinogalactan (AG) peptides, 21 (chimeric) fasciclin-like AGPs (FLAs), 5 other chimeric AGP, and 1 non-classical AGP), 57 EXTs (including 3 SP5 EXTs, 2 SP5/SP4 EXTs, 12 SP4 EXTs, 2 SP4/SP3 EXTs, 1
SP3 EXT, 11 "short" EXTs, 11 leucine-rich repeat-EXTs (LRXs), 11 proline-rich extensin-like receptor kinase (PERKs), and 4 chimeric EXTs, and 16 PRPs (including 11 PRPs and 5 chimeric PRPs), and 2 AGP/EXT hybrid HRGPs from the protein database derived from the completely sequenced Arabidopsis thaliana genome. We are currently exploring the use of Neural Networks and Hidden Markov Models to facilitate identification and classification of HRGPs. Preliminary results indicate that Hidden Markov Models show potential for further extending the set of AGPs.
Deciphering interaction networks forms the first step in annotating unknown proteins with functions, determining protein complexes, identifying target proteins and inventing new drugs. Over the past decade a large number of databases, websites and prediction tools have emerged that assign confidence scores to individual interactions. Unfortunately this growth in availability of interaction data is not coupled with increase in computational tools facilitating conversion of this data into useful information.

Molecular interaction networks are complex entities generally containing thousands of interactions. Prediction tools as well as experimental techniques aim at assigning quantitative metrics to each interaction edge within such networks. In most cases, the evidence about these interactions can be incomplete and associated with uncertainty. As a result, human judgment and expertise has to be exercised while deriving a set of high-confidence interactions after assessing each source of data. Methods are needed to examine this multi-dimensional, multi-source data in an automated and timely manner. The lack of computational tools facilitating such analysis for integration of evidence from multiple prediction and experimental sources such as Gene Neighborhood-GN, Gene Cluster-GC, Phylogenetic Profiles-PP, BIND (Bader et al.), and DIP (Xenarios et al.) is the motivation behind the CABIN software.
CABIN is developed as a plugin to Cytoscape (Shannon et al.) – an open source network visualization and analysis tool. CABIN promotes analytical reasoning for integrating evidence of interaction data from multiple sources by the use of interactive visual interfaces. Multiple coordinated views within CABIN fosters exploratory data analysis by users accommodating for expert domain knowledge. The functionalities available within CABIN maximize human perception and understanding of uncertain and complex data facilitating high quality human judgment with limited investment of the user’s time.
Smoothing spline mixed effects modeling of multifactorial gene expression profiles
Brandon J. Smith¹, Bruce R. Southey¹,², and Sandra L. Rodriguez-Zas¹,²,³,⁴
¹ Department of Animal Sciences, ² Department of Chemistry, ³ Department of Statistics, ⁴ Institute for Genomic Biology, University of Illinois at Urbana-Champaign, IL 61801, USA
{Sandra Rodriguez-Zas, rodrgzzs@uiuc.edu}

Abstract. Time-course gene expression data from microarray studies provide unique insights into the genes and pathways. The analysis of time-course microarray experiments is challenging because of the wide range of gene expression patterns, multiple sources of variation, and structure of the time-dependent measurements. A smoothing spline mixed model is proposed to describe complex time-course gene expression. A penalized likelihood-based approach was used to obtain estimates. The flexible approach was used to characterize the patterns of gene expression during honey bee behavioral maturation in a multifactorial experiment including time, race and host colony effect. Two spline dimensional bases were considered and a likelihood-based criterion was used to select a parsimonious spline basis that accommodates all the trajectories observed. Gene Ontology information of the microarray elements with significant differential expression across time within and across races and hosts indicated that, among others, genes that regulate development and synaptic communication are associated with maturation.

Keywords: Spline, Smoothing parameter, Mixed effects models, Microarray, Gene expression profile

1 Introduction

The characterization of gene expression profiles across time and other experimental conditions (i.e. environment) can provide insights into gene pathways and time-dependent functional or biological processes. The honey bee is a well-established model that exhibits marked behavioral maturation defined by time-dependent behavioral and genomic changes. Worker honey bees work in the hive for the first 2 to 3 weeks of adult life in a variety of tasks including brood care (nursing) and then shift to foraging for nectar and pollen for the remainder of their 4 to 6 week life [1]. Microarray studies across time have revealed extensive changes in brain gene expression associated with honey bee behavioral maturation [2], [3], [4], [5].

Gene expression profiles across time can be described using discrete age classes and analysis of variance. This approach is however, inappropriate to describe time-dependent phenomena because the data structure is ignored and each time point is often assumed to be an independent condition. Analysis of time-dependent observations using polynomial and spline functions can take into account the structure of the data and lead to more parsimonious and flexible descriptions of gene expression than factorial models. Polynomial regression functions of time have been
used to model changes in gene expression across time points in honey bee gene expression [3]. Likewise, simple spline models have been used to describe gene expression patterns across time [6]. The objective of this study was to assess the potential of spline mixed effects model to provide a flexible and parsimonious description of the trajectory of gene expression across time while accommodating technical and biological sources of variation.

2 Materials and Methods

2.1 Microarray data

The levels of gene expression of different behavioral maturation ages in the brains of honey bees from two races or genotypes (Apis mellifera mellifera and Apis mellifera ligustica) raised in one of two colonies (mellifera and ligustica), representing two environments, were measured in a cDNA microarray experiment [3], [4]. Full sisters were used to create the two host colonies for each race. Within each combination of race and colony, three nurse bees were sampled on days 0, 4, 8, 12 and 17 after adult emergence and three forager bees were sampled on day 17 after emergence. The gene expression from individual brains was assessed using the double-spotted Apis mellifera brain 9K version 3.0 cDNA microarray using the protocols [2], [4], [5]. Gene expression measurements were obtained from 108 microarrays in a loop design that maximized the direct comparisons between consecutive time points. Addition microarrays where used to interconnect honey bees across all the race-colony loops.

2.2 Data processing

The background subtracted fluorescence intensities were set to 1 when background was higher than foreground intensity and were log2-transformed. Spots that had poor quality or with low signal were removed from the data. After filtering, the intensities from the duplicated spots on the same microarray were combined into one value. The log2 intensity values were normalized using a linear-logarithmic transformation [7]. The adjusted gene expression measurements were analyzed using a two-stage model. The combined effect of microarray and dye effect was fitted in the first stage. Measurements from forager honey bees were removed from further analysis because only one forager time-point measurement was available. A spline mixed effects model was applied in the second stage to describe the trends of gene expression for each element using the R-language package Assist [8]. The model included the fixed effects of dye time, host, race, interaction effects and random effect of microarray. Multiple testing adjustment was incorporated by only deeming significant the model terms with the $P$-value$< 10^{-2}$. The Akaike likelihood-based criterion was used to identify the optimal dimensional basis to describe the expression profile of each element. Akaike information criterion indicated that the gene expression profile of
914 elements was better described using natural cubic than linear splines and there were no differences between the remaining elements for spline dimension. Therefore only a cubic spline was used.

3 Results and Discussion

A total of 7,638 elements were analyzed after filtering of weak and unreliable features or spots. After fitting the spline mixed effects model, there were 56, 2, 25, 10, 2, 8 and 37 elements that had significant differential expression across time, race, host, time×host, time×race, race×host and time×race×host, respectively. Figure 1 illustrates the flexibility and power of the smoothing spline mixed effects model proposed to describe microarray time-course data. For the element plotted in Figure 1a, host had a substantial effect on expression level and patterns because honey bees from any one race had different expression profiles across hosts. Figure 1b suggests that for the particular element under consideration, expression of the ligustica-mellifera and mellifera-ligustica honey bees were less variable across time than that from the same host-race combinations. The results uncovered using spline mixed effects models are consistent with studies that demonstrated the genomic plasticity of honey bees to changes in the environment [2]. Honey bees raised in predominantly mature colonies (high proportion of forager honey bees) tend to slow down their maturation process, meanwhile honey bees raised in predominantly young colonies (high proportion of nurse honey bees) tend to accelerate their maturation process. In this study, ligustica and mellifera honey bees are comparable to mature and young colonies respectively, because ligustica honey bees tend to mature faster than mellifera.

Fig. 1. Predicted gene expression profile across time for an element with significant time×race×host effect for two microarray elements. The lines denote the predicted trajectories for the race-host combinations mellifera-mellifera, mellifera-ligustica, ligustica-mellifera and ligustica-ligustica, respectively.

The incorporation of Gene Ontology information corresponding to the elements with significant differential expression across time (overall or within host or race levels) identified molecular functions and biological processes associated with the effects of maturation, genotype and environment. Genes that regulate synaptic growth at neuromuscular junction, development, formation and maturation of an ovum or
female gamete, nervous system development, mesoderm development, metamorphosis were observed to be differentially expressed across maturation age.

The spline mixed effect approach provided a smooth depiction of the overall gene expression trajectory and allowed for departures from the general trajectory due to host and race effects. The spline component provided a flexible description of gene expression trajectories and the linear mixed effects component accounted for technical and biological sources of variation typically observed in microarray studies. The capability to simultaneously model all sources of variation with a flexible approach enhanced the accuracy and precision of the estimates and predicted trajectories and the power of hypotheses tests.

Acknowledgements. Support of NIH/NIGMS grant 5R01GM068946 and NIH/NIDA grant 5P30DA018310-039003 are highly appreciated.

Reference

**DeltaProt: Molecular comparison of proteins based on sequence alignments.**

Steinar Thorvaldsen¹, Tor Flå¹ and Nils P. Willassen²

¹Dept of Mathematics and Statistics, Faculty of Science,  
²The Norwegian Structural Biology Centre,  
University of Tromsø,  
9037 Tromsø - Norway.  
steinart@math.uit.no

**Abstract.** In the toolbox DeltaProt we present statistical methods and trend-tests that are useful when the protein sequences in alignments can be divided into two or more groups based on known phenotypic traits such as preference of temperature, pH, salt concentration or pressure. The algorithms have been successfully applied in the research on extremophile organisms. We also provide procedures to plot the output from these tests for visualisations. DeltaProt is a Matlab© companion Toolbox that can be used freely for academic, non-profit purposes. Available from http://www.math.uit.no/bi/deltaprot/

1. Background

The development of DeltaProt has been motivated by the need to create a flexible software platform that enables easy statistical analyses of proteins with special preferences of temperature, pH, salt concentration or pressure [1-3]. For this reason, we have developed a Matlab toolbox as a special computational environment for molecular comparison. The reason we implemented such a toolbox in Matlab is to take advantage of the numerical computing capability of Matlab. We particularly work on identifying proteome-wide, and protein-specific, characteristics of cold adaptation (psycrophily). This is done by using comparative genomics on cold-adapted organisms, and similar genes from organisms with normal growth temperatures (mesophiles). In particular we have studied gamma-Proteobacteria from the order Vibrionales, where many genomes with different optimum growth temperature (T_{opt}) are already completed.

2. Statistical methods

The toolbox consists of a set of statistical routines with a variety of modelling functions. We consider both the amino acid sequence compositions, and the substitution patterns, to determine whether there are underlying trends that explain the observed variation between the phenotypic groups to be analysed. More than 80 different physicochemical properties [4] of the amino acid may also be applied in order to reduce the sequence alphabet to measurements. Each situation is analysed by appropriate statistical methods:
• **Composition:**
  Linear regression

• **Substitutions:**
  Fisher’s exact test, Chi-square tests, Mantel-Haenszel test

• **Properties:**
  Wilcoxon paired test,
  Non-parametric regression based on Mann-Kendall statistics

If the secondary and/or 3D structure is known, or may be predicted, the analyses can be performed in each of these regions.

When alignment data consists of species from different taxonomic orders, this information may be utilised to model a *stratified two-way design* [3]. In a stratified design (also called blocked analysis or matched analysis), the data are selected from two or more strata that are formed from important covariates such as taxonomic order. A stratified two-way design, where also taxonomic order is taken into consideration, yields a better statistical analysis. This model is more robust against differences due to differences between different phylogenetic lineages, and hence appropriate to detect possible trends independent of lineage. Stratification of the data into disjoint groups increases the power of the tests.

The statistical approach may assume pairwise *independence* among the sequences samples in each group at each strata, or it may be modified to treat phylogenetically *dependent* sequences within the groups. When there are multiple dependent samples in a group, we compute the mean values in each group at each stratum, and apply the statistical test to the means as representative observations. This use of group sample mean also removes some of the random genetic drift (“noise”).

When multiple hypothesis tests are carried out, significance levels should be adjusted to account for the increased probability of false positives. For multiple test correction we use Benjamini and Hochberg False discovery rate analysis (FDR) [5].

3. Some visual results

![Graph](image)

**Fig. 1.** Composition of amino acids in ortholog sequences from three groups of bacterial Endonuclease I, ranked according to frequencies in the Non-marine group.
Fig. 2. Visualization of the number of pairwise substitutions observed in a comparison of 298 ortholog proteins between two groups. The size and color of each marker indicates the magnitude of the substitution (see color-bar). Favored substitutions (P-value < 0.05) in the Mesophile→Psychrophile direction are marked with upward-pointing triangles, and the non-favoured substitutions (P-value < 0.05) are marked with downward-pointing triangles.

Fig. 3. Peptide plot of a physicochemical property (helix propensity) for three groups of Endonuclease I from different environments. A box filter of size m x 3 was used as smoothing technique to recover the underlying structure in the data, where m is the number of sequences in the group. There is a significantly higher helix propensity in the salt adapted group (P-value=0.00001)
The program suite consists of two main program scripts and 30 procedures. The main scripts should be modified to provide access to new alignment data. Figure 1 and 2 shows examples of visual output when analysing proteins from marine, brackish water and fresh water environments, while Figure 2 shows all substitutions between two temperature defined groups. A tilde (\textasciitilde) indicates deletion. Routines for comparing DNA molecules at the nucleotide level will be added in future development.

Acknowledgements
The present study was supported by the National Program for Research in Functional Genomics in Norway (FUGE), and the sequence alignment of endonuclease I was kindly provided by Bjørn Altermark.

References
Using Logical Sets to Target Gene Expression Patterns

Timothy L. Tickle\textsuperscript{1} and M. Taghi Mostafavi\textsuperscript{1}

College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte NC 28223, USA

1 Introduction

Recent advances in molecular biology have enabled researchers to begin to address the genetic and proteomic complexity associated with different disease states. One of these technologies, microarrays, has become prevalent in gene expression quantification. Sets of microarrays are used to measure patterns of genetic expression throughout treatment groups for each of thousands of genes in an experiment. Researchers use many types of algorithms to select genes in an experiment, targeting differentially expressed patterns of gene expression coherent to their experimental model. Many algorithms approach microarray analysis with optimizing algorithms using random features which can give different results for the same data raising an issue of reproducibility. A more quantitative approach commonly used is to perform a variant of a t-test or ANOVA. These measurements rely, however, on a central tendency measurement for the group and not the individual data itself. This paper takes a computational approach to selecting significant genes by combining exhaustive testing with logical set theory in order to target genes that exhibit specific patterns of expression through the experiment microarrays.

2 Methods

Fifty-six ovarian cancer related samples were processed with Affymetrix Human Focus Gene Expression Chips using the manufacturer’s recommended protocols. Samples were collected in cooperation with the Blumenthal Cancer Center at Carolinas Medical Center (CMC) and processed at the Molecular Biology Core Facility (CMC). Based on pathology, samples were divided into 3 different groups (Normal, Benign, and Malignant). After Affymetrix Microarray Analysis Suite 5.0 (MAS5) calculated signals of gene probe sets, all microarray data was quantile normalized [1]. After normalization, signal was checked using M vs A plots generated by dChip.

A baseline normal sample was selected by pathology report and comparative metrics were computed against the selected baseline for all other microarrays using the MAS5 SDK implementation of the MAS5. Change p-values (a measure indicating the significance and direction of change of a gene instance from a common baseline) were used in exhaustive comparisons between microarray
groupings to identify significant genes. The bioconductor package multtest for R was used to convert the change p-values to false discovery rates [2]. The “BY” false discovery rate algorithm was used at a rate of .05 [3].

3 Algorithm

The experimental model used for this algorithm uses one control group (referred to as the “Normal” group) and two test groups (referred to as “Benign” and “Malignant”). This is a commonly used experimental model (targeted for cancer studies) and can be extended to other models. Genes are targeted that are not considered changing in the normal sample group from a designated baseline microarray, but are at least changing in one of the test groups. The first three columns of Table 1 lists all targeted gene expression patterns.

When comparing p-values of genes in different sample groups, p-values fulfilling two different types of logical criteria were targeted. These criteria were based on the fact that a p-value will exist as one of the three situations (given a threshold \( \alpha \)): 1. \( p-value \leq \alpha \), 2. \( p-value \geq (1-\alpha) \), and 3. \( \alpha < p-value < (1-\alpha) \). The first criteria selects a gene if \( \alpha < F < (1-\alpha) \) and \( S \leq \alpha \), where F is an instance of the gene in one group (or a first group) while S an instance of the same gene in another group (or a second group). This would indicate a gene with significant increase from the baseline in the first group and no significant change from the baseline in the second group. The next criteria selects genes if \( \alpha < F < (1-\alpha) \) and \( S \geq (1-\alpha) \), indicating no change in the first sample group but a decrease in the second.

To target a specific pattern of gene signal expression throughout an experiment of more than two groups (in this example three categories) a combination of comparison were needed (Table 1). Multiple comparison sets were needed since when selecting for a criteria for two groups, for instance Normal vs Benign, the third group (Malignant) is not held under any constraints.

Since any gene exhibiting significant change in the correct expression pattern was selected, genes with only one occurrence of the correct expression pattern are initially retained. To reduce the genes without consistent expression patterns, genes are required to have in each group more than 51% agreement among p-values being in one of the three previously mentioned p-value groups.

4 Results

The most commonly found patterns were No Change, No Change, Increase and No Change, No Change, Decrease. These patterns share the trend of both the first and second groups (normal and benign) not changing from the original normal baseline in the t-tests while the third group (malignant) changing in their respective directions. These patterns were significantly more frequent than any of the other patterns (Table 2). To investigate if this trend is due to data or the algorithm, Principal Component Analysis (PCA) on the samples, using the R statistical package, was performed. PCA showed a strong trend for normal
Table 1. The table below summarizes the use of logical criteria to target expression patterns. The first three columns (Normal, Benign, Malignant) show the targeted change in expression in each group of samples and the summary of logical criteria is the combination of logical criteria and comparison sets indicated in table.

<table>
<thead>
<tr>
<th>Gene Expression Pattern</th>
<th>Normal</th>
<th>Benign</th>
<th>Malignant</th>
<th>Summary of Logical Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Change Decrease No Change</td>
<td>{\alpha - 1 &gt; N &gt; \alpha} \cap {B \geq (1 - \alpha)}\cap {\alpha - 1 &gt; M &gt; \alpha} \cap {M \geq (1 - \alpha)}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Change No Change Decrease</td>
<td>{\alpha - 1 &gt; B &gt; \alpha} \cap {M \geq (1 - \alpha)}\cap {\alpha - 1 &gt; N &gt; \alpha} \cap {M \geq (1 - \alpha)}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Change Decrease Decrease</td>
<td>{\alpha - 1 &gt; N &gt; \alpha} \cap {B \geq (1 - \alpha)}\cap {\alpha - 1 &gt; N &gt; \alpha} \cap {M \geq (1 - \alpha)}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Change Increase No Change</td>
<td>{\alpha - 1 &gt; N &gt; \alpha} \cap {B \leq \alpha}\cap {\alpha - 1 &gt; M &gt; \alpha} \cap {B \leq \alpha}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Change No Change Increase</td>
<td>{\alpha - 1 &gt; B &gt; \alpha} \cap {M \leq \alpha}\cap {\alpha - 1 &gt; N &gt; \alpha} \cap {M \leq \alpha}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Change Increase Increase</td>
<td>{\alpha - 1 &gt; N &gt; \alpha} \cap {B \leq \alpha}\cap {\alpha - 1 &gt; N &gt; \alpha} \cap {M \leq \alpha}</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and benign samples to cluster together and for malignant samples to be separate from normal and benign samples (data not shown). This supports the trend of having more significant genes only changing in the malignant samples as a valid trend in the data rather than an effect of the analysis.

Table 2. Expected occurrences refers to the number of pattern instances found that were found with the expected combination of logical criteria and comparisons. Occurrences not expected refers to pattern instances that occurred with logical criteria and comparisons not expected to target the pattern.

<table>
<thead>
<tr>
<th>Expression Pattern</th>
<th>Expected Occurrences</th>
<th>Unexpected Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Change, No Change, Increase</td>
<td>288</td>
<td>0</td>
</tr>
<tr>
<td>No Change, Increase, Increase</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>No Change, Increase, No Change</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>No Change, No Change, Decrease</td>
<td>354</td>
<td>0</td>
</tr>
<tr>
<td>No Change, Decrease, Decrease</td>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>No Change, Decrease, No Change</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Other genes were found as well. These genes involve patterns that show a difference in change between the first and second groups (normal and benign). These types of significant genes were found less frequently. This is also corroborated by the PCA analysis.
One of the genes, PAX8 was selected to be biologically validated. Using real-time RT-PCR and histochemistry, PAX8 was shown to have a general association with malignant ovarian cancer both in primary tumors and cell lines [4]. Other genes found by this technique are known to be associated with cancer, some specifically with ovarian cancer including ERBB2 (Her2/neu) [5], NME1 [6], Ras family genes (H-Ras, K-Ras and N-Ras) [7], VEGF [8], and TGF-β[9]. Other genes selected have been shown to be associated with cancer.

Acknowledgments We would like to acknowledge the Blumenthal Cancer Center at Carolinas Medical Center (CMC) for providing human ovarian samples and their continued support of this work. We would also like to acknowledge the Molecular Biology Core Facility at CMC for sample preparation and raw data collection.

References

Database for Structural Analysis of HIV Protease

Yunfeng Tie¹, Hao Wang², Robert W. Harrison¹,², and Irene T. Weber¹

¹ Department of Biology, Georgia State University, Atlanta, Georgia 30303, USA
² Department of Computer Science, Georgia State University, Atlanta, Georgia 30303, USA.
{Yunfeng Tie, Hao Wang, Robert W. Harrison, Irene T. Weber, iweber@gsu.edu}

Abstract. A prototype domain-specific database has been designed to organize and automate the comparative analysis of protein structures. The database is designed to provide the basic functions of structural analysis with a flexible modular computing architecture. Structure data are represented by a hierarchical arrangement of objects for atom, residue, chain and PDB. The current implementation includes functions for superimposing up to 8 structures on a reference structure, and calculation of user-selected interatomic distances. This database has been tested with crystal structures of HIV protease mutants in complex with various inhibitors. Antiviral HIV protease inhibitors developed with the aid of structures are important drugs to treat HIV/AIDS. This database for comparison of structures has the potential for further development for improved analysis of HIV protease mutant structures, drug design, and extension to structural analysis of other proteins.

Keywords: Domain specific databases, domain specific queries, HIV protease, structural analysis.

1 Introduction

Comparative analysis of protein structures is critical for understanding protein function and the relationships among different proteins, engineering of proteins with new functions, and structure-based drug design. A prototype domain specific database has been developed for comparison and analysis of structures of HIV protease with inhibitors. HIV protease is one of the most effective drug targets for HIV/AIDS. Therefore, hundreds of crystal structures of HIV protease-inhibitor complexes have been determined and used to guide the development of antiviral inhibitors. Indeed, these efforts illustrate the enormous potential of rational structure-based drug design [1]. Nine protease inhibitors are used in Highly Active Anti-retroviral Therapy, which extends the life of AIDS patients. However, the long-term effectiveness of antiviral treatment is severely limited by drug resistance and narrow choices of available drugs.

Few databases have been developed for HIV. Most databases summarize valuable information about sequence, resistance, immunology, and vaccine trials, rather than structural data. Online databases include the Los Alamos HIV database at http://hiv-web.lanl.gov/content/index, Stanford HIV sequence and resistance database at http://hivdb.stanford.edu and the National Institute of Allergy and Infectious Disease site of anti-HIV compounds at http://www.niaid.nih.gov/daptdpdb [2]. One database containing about 200 structures of HIV protease-inhibitor complexes is
available at the National Institute of Standards and Technology (NIST) at
http://srdata.nist.gov/hivdb [3]. The NIST database provides a tree of fragments for
each inhibitor and has an analysis tool to perform fragment searches for inhibitors
containing similar fragments [4]. However, despite these strengths, this database
system has deficits. It is a website based program, and is not designed for local lab
use. It cannot be used securely with unpublished and proprietary data. Most seriously,
it lacks the tools for analyzing the detailed interactions between atoms and chemical
groups in different complexes. Therefore, we have developed a prototype database to
efficiently analyze and compare HIV protease-inhibitor structures and interactions in
order to explore new drug-design principles. This prototype database can be expanded
to organize, compare, and analyze any set of related protein structures.

2 Design

The system design focuses on how to efficiently make domain specific queries about
protein structure. Protein structure is not a primitive data type in an SQL database.
Therefore, the data require conversion into and out of an efficiently searchable
representation. A domain-specific user friendly and context sensitive interface is used
to facilitate this conversion.

2.1 Database

About 240 HIV protease structures are available in the PDB [5], including more than
50 crystal structures of different mutants of HIV protease from our group. Many of
our unpublished structures contain confidential information. Therefore, organizing
our data was the primary motivation for developing the database, although other data
have been included from the PDB. The database is designed to provide the basic
functions of structural analysis with a powerful and flexible computing architecture.
The PDB ID is not sufficient to be a primary key. Therefore, a primary key is
generated by concatenating C for category, P for protein, I for inhibitor, K for PDB
ID, S for space group and Y for the year of publication.

The structural data must be parsed and extracted into a set of programming objects
in a hierarchical layout. The PDB object instance is saved as a BLOB type [6]: The
HIV protease data are first processed and wrapped into an object instance and then,
the object instance is saved into a BLOB field of the HIV data table. The retrieval of
data similarly takes two steps: the data are located and retrieved as a BLOB object
BLOB object and then recast into a PDB object.

2.2 Analysis Functions

Data Searching, Synchronization and View. The structural data are named by
feature composition so one level of filtering can be performed by parsing the data
name. Once potential targets are defined, the system automatically synchronizes the
target data between backend database and the client. Target data are retrieved from
the remote database, and wrapped into individual objects locally.
The display of structural data follows the Model-Control-View strategy. All the structural data are displayed as a tree. It is possible to view by category, protein and inhibitor. In each mode, the individual PDB object will appear in the third layer of the display tree with a link to text or graphic display with a program like Rasmol [7].

**Interatomic Distance Calculation.** The interactions between protease and inhibitor are characterized by the type of atom and the distance between atoms. Two atoms with separation of about 3.5 to 4 Å are considered to form good van der Waals interactions, while separations of 2.6 to 3.2 Å distance between a hydrogen bond donor atom and a hydrogen bond acceptor atom are defined as hydrogen bonds. Once an object is selected, the program will automatically go to the lower level leaf of atom object and take all the atoms within that object. Distances are measured between every pair of atoms in the two chosen objects. A group calculation allows the user to input two atom objects within one PDB object, and then put user-selected PDB objects into the group. The system retrieves the types and positions of the two input atoms, and applies the calculation to the equivalent atoms for all input PDB objects. The results are filtered with user-specified limits, and exported in EXCEL format.

**Structure Superimposition and Visualization.** The system calls the FUD program, which implements the algorithm of Hermans & Ferro [8] to find the least square error coordinate superimposition of two protein structures. A maximum of 8 HIV protease structures can be selected and superimposed on the same reference file at the same time. The statistics and superimposed structure are generated and stored for each pair of structures. A graph shows the differences for superimposed Cα atoms for each residue, and the global root mean square deviation for each pair of superimposed structures.

**Inhibitor Compound Search.** The chemical compound search engine developed in our group [6, 9] is interfaced with this database so that inhibitor fragments can be used to search for similar compounds.

### 3 Implementation

The system comprises the back-end database and the front-end analysis agent program. The backend is implemented with the Oracle database engine. Structural and inhibitor data are saved in the database. The domain specific user interface is implemented in a front-end analysis agent, written in Java. Some analysis functions, such as the distance calculation, are implemented directly in this program. Other functions call external programs, which are wrapped by a generic daemon [9]. The wrapper daemon significantly decreases the coupling between the function call and external program, which facilitates future modifications.
4 Conclusion

A prototype domain specific database has been developed using Oracle for organizing and analyzing protein structure data. The structure database was designed for comparison of protein structures in a local setting to ensure security for confidential information. The database system was evaluated with data for HIV protease-inhibitor crystal structures used in structure-guided drug design for HIV/AIDS therapy. The database contains both published data and proprietary data for novel inhibitors. Most common tasks of data processing and domain specific calculations are performed in a user friendly and context sensitive working environment. The currently implemented analysis functions are superimposition of structures and interatomic distance calculations. The structure database is coupled with a chemical search engine. Application of this structure database has improved the routine analysis for HIV protease-inhibitor complexes, enabled the extraction of hidden patterns, and the ability to explore new drug-design principles. The design incorporates a flexible architecture, with a defined wrapping mechanism for “plugging in” applications, so that other calculations or analytical programs can be easily incorporated. It has the potential for further development for structure-guided drug design, and extension to structural analysis of other proteins.

References

A integrated solution based on distance method for reconstructing phylogenetic trees

Cristianno Vieira, Glauber Gonçalves and Martha Torres

Departamento de Ciências Exatas e Tecnológicas da Universidade Estadual de Santa Cruz, Ilhéus, Bahia, Brazil

Abstract. This poster presents DiGrafu that is a solution based on distance method, this solution integrates FastME, Weighbor, BIONJ and NJ programs. Besides, this poster presents validation of proposed solutions.

1 INTRODUCTION

There are many phylogenetic tree reconstruction programs [1] [2] [3] [4] [5]. Therefore, it is difficult for user to choose which program to use. In order to alleviate this problem, many researchers have realized studies for evaluating the developed programs [6] [7] [8] [9].

The main existing methods for reconstructing phylogenetic trees are: maximum likelihood, parsimony and distance. Although, the distance methods are the ones that present the lesser exactness, they are not less important because, on the contrary to other methods, its performance allows to reconstruct phylogenies with a very great number of species. This poster presents a solution based on the distance method that integrates NJ [17], BIONJ [1], Weighbor [11] and FastME [18] programs, providing to the user the solution that better adjusts to his necessities of execution time and/or exactness.

The remainder of this poster is organized as follows: In next section we describe DiGrafu. In section 3 we explain the methodology os this work. In section 4 we report the results. Finally, in section 5 we present conclusions and a description of future work.

2 DiGrafu

There are many programs based in distance method, we studied and selected five important programs: NJ, BIONJ, UPGMA, Weighbor and FastME. Then, we developed a tool that allows to explore the best characteristics of these programs. Each program uses as input a distance matrix which is calculated through Dnadist 3.6 [12] program. This program is also integrated in DiGrafu.

Using Digrafu, the user does not worry about choosing the program to be used. Our program selects the best option for him. In order to do that, Digrafu’s choice is based on the input data and an additional option provided by the user. In the last option, the user can choose if his priority is execution time, accuracy or a mix of both.
3 Methodology

We used synthetic data sets in order to evaluate the distance-based methods under consideration. In a simulated environment, the true tree is artificially generated and forms the basis for comparison among phylogenetic methods. We use the 5,000 trees generated by Guindon and Gascuel in [6], each tree has 40 taxa. These phylogenies have a broad variety of deviations from the molecular clock and various evolutionary rates. The branch length mean is equal to 0.06 substitutions/site and the mean of the ratio of the length of the longest to the length of the shortest lineages is equal to 3.4. According to [6], these values come from an analysis of substitution rates in various organisms and of numerous published phylogenies. Then, we generated sequences 500 base pair (bp) in length from these phylogenies using Seq-Gen [13]. This program generates sequences data sets based in a model tree and a evolutionary model and it is very utilized for generating synthetic data sets [6] [16] [8]. We generated sequences under the Kimura 2-parameter (K2P) model [14], with a transition/transversion ratio of 2.0 and the Jukes Cantor (JC69) model.

These sequences are given to the phylogenetic reconstruction methods, which infer a tree based on the given sequences. The inferred trees are then compared against the model tree for topological accuracy. We use the Robinson-Foulds distance [15] to measure the error between trees. This distance corresponds to the number of internal branches that are found in one tree and not in the other one. The Robinson-Foulds distance is calculated using the TreeDist program from Phylip [12]. This value is not normalized, if it is 0.0 then both topologies are identical, therefore, it means that when the distances are smaller, then the trees are more similar.

We plotted the Robinson-Foulds distance against the maximum pairwise divergence (MD) in the synthetic data sets. The (uncorrected) divergence between two sequences is the proportion of sites where both sequences differ. The figure 2 shows the results for K2P and JC69 models.

The FastME solution has four versions: the GME+FASTNNI solution is the combination of greedy ordinary least-squares minimum evolution tree construction algorithm (GME) and the FASTNNI tree swapping algorithm. The GME+BNNI solution is the combination of GME with the BNNI tree swapping algorithm based in the balanced minimum evolution framework. The BME+FASTNNI solution is the combination of balanced minimum evolution tree construction algorithm (BME) and the FASTNNI algorithm. The BME+BNNI solution is the combination of BME with BNNI.

The results are in accordance with expectations and with previously published simulations [6] [7] [8]. When the MD is low, phylogeny reconstruction is hard because there is not enough information in the data to estimate the short internal edges. With a high MD, saturation corrupts the phylogenetic signal and reconstruction is again hard. This explains why all methods perform better with medium divergences rates. The figure indicates that UPGMA presented the worse values (higher values). This method is old and it is not good for sequences
Fig. 1. Topological accuracy of Neighbor, UPGMA, BIONJ, Neighbor, GME+FASTNNI, GME+BNNI, BME+FASTNNI and BME+BNNI as a function of the divergence between sequences.

Also, we plotted the execution time against the sequence divergence. This execution time was measured using the times library of PERL and it includes the execution time of Dnadist program. The Figure 3 shows the results. The figure points that the methods had a similar behavior, except Neighbor that for using likelihood calculations to develop the exactness, increases the computational requirement in considerable way. BIONJ, always was remained as the method fastest, except in the point of MD 80% of K2P model, where the versions of FastME had a punctuation better.

### 4 Results and Discussion

Thus, DiGrafu is implemented taking into consideration the previous results. Where, UPGMA is not taken in account by its poor performance. DiGrafu chooses the best method in each value of sequence divergence and for each model. When the user selects the execution time as option then DiGrafu executes BIONJ, which is more efficient in almost all cases. When the user selects the execution time and accuracy, DiGrafu executes the method that provides to greater exactness, except Neighbor. The FastME method, in the versions BME+BNNI and GME+BNNI, was most used in this option. Neighbor is ignored because its execution time is very high compared with the other methods. When the user selects the accuracy, then DiGrafu chooses the method with bigger exactness. Neighbor is the method most used and FastME, in the versions BME+BNNI and GME+BNNI, is chosen few times.

For validation of the DiGrafu method, this was submitted to the same previous tests of performance. The results are in accordance with expectations. The result of DiGrafu[ta] (execution time and accuracy) is always more accurate or equal
Fig. 2. Execution time of Neighbor, UPGMA, BIONJ, Weightor, GME+FASTNNI, GME+BNNI, BME+FASTNNI and BME+BNNI as a function of the divergence between sequences.

to DiGrafu[t] (execution time) and DiGrafu[a] (accuracy) was most accurate of the three. DiGrafu[t] was the fastest and DiGrafu[a] was most demanding in processing terms. Digrafu is a program developed in PERL, it integrates the programs above mentioned.

5 Conclusions

We have developed, implemented and tested DiGrafu. Our solution includes four popular distance-based programs; it selects which program to use analyzing input data set and efficiency and/or accuracy criteria. In future works, we intend to develop other integrated solutions for maximum likelihood, parsimony and Bayes methods.

Acknowledgment We would like to thank FAPESB and UESC by the grants, as well as FAPESB and LABBI for the infrastructure.
References


67

ISBRA 2007 Poster Proceedings
A Survey of Basecalling Algorithms

Andy Ju An Wang
Southern Polytechnic State University
1100 South Marietta Parkway, Marietta, GA 30060, USA
jwang@spsu.edu

Abstract. Basecalling is critical for human genetics and bioinformatics research. The term, definition, and process of basecalling have been evolving over time with a variety of standards and technologies coexisted in the literature. It can be difficult for beginners to navigate and understand the most important aspects of basecalling and associated algorithms. This paper provides a brief overview of basecalling algorithms for computer scientists who are interested in bioinformatics research and development. The major basecalling algorithms discussed here include ABACUS, Affymetrix GDAS, and Model-P.

Keywords: DNA sequence, basecalling algorithms, bioinformatics.

1 Introduction

A focusing theme in molecular biology and bioinformatics is to discover the DNA sequences that contribute to particular cell functions of interests, such as human phenotypes, or particular diseases. Genes are the fundamental units of DNA functions. Though all human have the same genes, some of these genes contain sequence differences that make each person unique. Therefore, DNA sequences can be used to detect inherited diseases and catch criminals, among other useful applications. Basecalling is a process to determine DNA sequences. It is done through manual, semi-automated, or complete automated processes.

Bioinformatics involves the manipulation, searching, and data-mining of DNA sequence data. A sequence alignment is a way of arranging the primary sequences of DNA to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between sequences. The building blocks of DNA is the nucleotide. Each nucleotide contains one base, either A (adenine), T (thymine), G (guanine), or C (cytosine). DNA sequencing is a laboratory technique for determining the nucleotide “sequence” or arrangement of A, C, G, T nucleotides in a segment of DNA. Through the Human Genome Project (HGP), the high-quality reference DNA sequence for the human genome’s 3 billion base pairs was completed in April 2003. However, basecalling is still vital in research because we need DNA resequencing to find out those particular genes associated with numerous diseases and disorders. DNA resequencing is the task of sequencing a DNA region for an individual given that a reference sequence for this region is already available for the specific species.

Basecalling is critical for human genetics and bioinformatics research. The term, definition, and process of basecalling have been evolving over time with a variety of standards and technologies coexisted in the literature. It can be difficult for beginners
to navigate and understand the most important aspects of basecalling and the associated algorithms. This paper provides a brief overview of basecalling algorithms for computer scientists who are interested in bioinformatics research and development. In the following, we will first discuss different methods for basecalling in the next section. Then we present a number of basecalling algorithms that have been developed. A brief discussion and comment about basecalling algorithm design and implementation are presented at the end of this paper.

2 Different Methods for Basecalling

There are many different methods for basecalling. For the purpose of this paper, we discuss three types of methods: manual methods, computer-aided methods, and DNA microarray methods.

2.1 Manual Methods

A typical manual method for basecalling is called chain-termination, or Sanger method, as it was developed by Fredrick Sanger in 1977 [5]. This method starts with a DNA sample of interest and an oligonucleotide primer complementary to a specific site on the template strand. For each of the four bases (A, C, G, T), a reaction is carried out in which DNA polymerase synthesizes a population of labeled single-stranded fragments of varying lengths, each of which is complementary to a segment of the template strand and extends from the primer to an occurrence of that base. Autoradiography is used to detect the radioactive sequencing fragments. The sequence determined from the autoradiogram is complementary to the sequence on the template strand, so the base sequence of the template is inferred.

The manual method can only be used to sequence approximately 200 – 400 nucleotides in a single reaction. For a large-scale sequencing, we need a computer-aided approach.

2.2 Computer-Aided Methods

Smith et. al. [4] developed an effective basecalling method with the help of computer automation. In this method, the DNA fragments are labeled with fluorescent dyes attached either to the primer or to the dideoxy chain-terminating nucleotide. Typically a different dye is used for each of the four reactions, so that they can be combined and run in a single gel lane. A laser beam is used to stimulate the fluorescent dyes creating different color patterns for each nucleotide. The fluorescent pattern is processed and converted by a computer to reveal the DNA sequence of bases.

2.3 DNA Microarray Methods

Microarrays are parallel devices enabling researchers to test all the genes expressed in a tissue very quickly. A single microarray can have 300,000 spots of DNA, each containing unique sequences of DNA for a different gene.

Researchers are also using microarrays to compare patterns of expressed genes in tissues under different conditions. For instance, cancer cells can be compared to normal cells to look for genes that may be involved in cancer formation. High-density VDAs (Variation Detection Arrays) manufactured by Affymetrix have been widely used in DNA sequencing. Each VDA has 300,000 features with a feature size of 24 by 20 microns. A feature consists of one million copies of a 25-bp long oligonucleotide
probe of defined sequence. To query a specific site determined from the human genome reference sequence, four features are tiled on the VDA. The four features differ only at the central or 13th base, which consists of each of the four possible nucleotides. Each human genome site is queried for both the forward and reverse strands at different location on the VDA.

A single DNA microarray can contain thousands of genes. Researchers can use microarrays to screen a patient for a pattern of genes that might be expressed in a particular disease condition.

3 Basecalling Algorithms

The goal of basecalling software is to produce a DNA sequence as accurate as possible. In this section, we discuss five basecalling software products and their basecalling algorithms.

3.1 ABI Basecalling Algorithm

ABI basecalling algorithm [6] uses mobility curves to predict the peak spacing, and identifies the most likely peak in intervals of the normal peak spacing, assigning an N in the absence of a good choice. Therefore it adds and removes bases using a criterion involving the uniformity of peak spacing.

3.2 Phred Basecalling Algorithm

The phred basecalling algorithm [9, 10] uses a four-phase procedure to determine a sequence of basecalls from the processed trace. In the first phase, idealized peak locations (predicted peaks) are determined. In the second phase, observed peaks are identified in the trace. In the third phase, observed peaks are matched to the predicted peak locations, omitting some peaks and splitting others; as each observed peak comes from a specific array and is thus associated with one of the four bases, the ordered list of matched observed peaks determines a base sequence for the trace. In the final phase, the uncalled observed peaks are checked for any peak that appears to represent a base but could not be assigned to a predicted peak in the third phase, and if found, the corresponding base is inserted into the read sequence.

3.3 ABACUS

ABACUS (Adaptive Background genotype Calling Scheme) [7] is an automated statistical system for DNA basecalling. It can be applied in experiments in which the target DNA sequences are either haploid or diploid. ABACUS assumes that the observed fluorescence intensities are normally distributed within features. Each feature consists of one million distinct oligonucleotides of identical composition. A series of statistical models are developed under the assumption of the presence or absence of various genotypes in the target sample. The likelihood of each statistical model for a given genotype is calculated independently for both the forward and reverse strands and is combined for the overall likelihood of the model. A “quality score” is assigned to each VDA genotype. A site genotype is “called” when one model fits the data sufficiently better than all other models. The average call rate was reported to be > 80%, and the call accuracy was 99.9999%.
3.4 Affymetrix Basecalling Algorithm
The GeneChip® DNA Analysis Software (GDAS) implemented a basecalling algorithm based on [7]. The average call rate was claimed to be 95.4%, and the automatic sequence call accuracy was 99.998%.

3.5 Model-P Basecalling Algorithm
Zhan [2] proposed a basecalling method called Model-P, using physical models based on the sequences of the probe and the target to predict feature intensities in resequencing microarrays. Model-P takes into consideration the expected feature intensities for different potential genotypes. It was claimed that Model-P had better performance at high call rates compared with ABACUS on a test dataset and on relatively AT-rich regions.

4 Discussion and Conclusion
Genes have been pinpointed and particular sequences in those genes associated with numerous diseases and disorders including breast cancer, muscle diseases, deafness, and blindness [8]. Basecalling algorithms play an important role in DNA sequencing and resequencing. These algorithms should be studied and compared in terms of call rate, accuracy, reliability, and performance. Base call rate is defined as the fraction of individual bases that can be clearly detected and identified by an algorithm. Base call accuracy is defined as the fraction of correctly classified bases among a set of possible positions and samples.

While microarrays are becoming invaluable for biologists studying how genes interact with each other and understanding disease mechanisms, basecalling is a critical step of the analysis of DNA resequencing microarray data for single nucleotide polymorphism discovery and genotyping [2]. This is an interesting area for computer scientists to contribute to the field.

References
6. ABI PRISM, DNA sequencing analysis software, user’s manual. PE Applied Biosystems, Foster City, CA, 1996.
Ordered Combinatorial Feature Selection: An Information Portal for Multiple Indexing Sequence Alignment

Hsin-Wei Wang¹, Jian-Ming Chen¹, Wei-Yao Chou², Margaret Dah-Tsyr Chang², and Tun-Wen Pai¹∗

¹Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, Republic of China
²Institute of Molecular and Cellular Biology & Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan, Republic of China
∗twp@mail.ntou.edu.tw

Abstract. An ordered combinatorial feature database is designed for identification of sequential consensus motifs as important functional and/or structurally critical peptides for all protein families defined by Pfam. The ordered combinatorial features in the proposed database are extracted from each protein family employing a multiple indexing sequence alignment that performs interval jumping searching algorithms and center-star alignments. The interval jumping searching algorithm is designed to find consensus motifs by using the combination of hashing encoding, quick sorting and interval jumping techniques, which provides an alternative manner to achieve approximate matching functions in linear time. The center-star alignment identifies combinatorial features among the query sequences, and the hierarchical clustering algorithms combining with bitwise comparison operations perform exclusive group feature extraction for each individual subgroup. In this research, all protein families defined by the Pfam database are pre-analyzed by the proposed algorithm, and their respective combinatorial group features are collected in a database for further referencing and applications. To verify the biological meanings of the searched patterns, each feature was scanned by Prosite (release 20.4), and a hit rate of more than 73.6% was achieved for frequently occurring signatures.

Keywords: combinatorial feature, exclusive group feature, multiple indexing sequence alignment, protein family.

Availability: The ordered combinatorial feature database is available at http://misa.cs.ntou.edu.tw/MISA.Database, and the details of the proposed algorithm can be referred to the documents as described previously (1,2)

1 Introduction

The huge quantities of genome projects being carried out worldwide have led to an explosion of sequence data in public databases. To establish functional and
evolutionary similarities among a set of protein sequences, multiple sequence alignment has been considered as the intuitive methodology in the characterization, classification, and correlation of biological functions. Therefore, several algorithms have been proposed to perform a multiple sequence alignment (MSA) for inferring sequence homology and phylogenetic relationships. To achieve these goals, various algorithms have been developed to execute sequence alignment and to extract important features (3). Most state-of-the-art algorithms employ progressive strategies by extending the results of pairwise alignment to three or more sequences; sequences are hierarchically aligned in pairs according to a pre-calculated tree structure based on mutual sequence similarities. However, these methods can sometimes result in misalignment and will discard common sequence motifs because of errors made in previous steps. To avoid these problems, other strategies are necessary, such as those that resemble the manual multiple alignment performed by biologists based on restricted selection of similar block candidates. This methodology is more suitable for allocating functional motifs than are conventional multiple alignment algorithms. It starts by searching putative functional motifs or homologous segments, followed by defining the common factors as constrained positions, and finally performing intermediate sequence alignment. In general, there are two types of computational segment-based methods. One is exact complete block, which searches only identical matched segments or tolerant segments, i.e., all segments that occurred in all query sequences. The other is incomplete block, which considers the anchored segments not shared by all query sequences (4,5). After identification of the blocks of interest, the following procedures achieve piecewise alignment to carry out multiple sequence alignment through partial selected blocks, and the performance is strongly based on various definitions of score functions (6). Because of the ability to emphasize local characteristics, the results of the computational segment-based method for multiple sequence alignment often provide more information in the context of biological significance. Thus, we propose an alternative segment-based methodology that performs interval jumping searching and center-star alignment to enhance the efficiencies and effectiveness of multiple sequence alignment. In addition to propose the novel MSA algorithms, we have performed the developed system on Pfam database and created a combinatorial feature database for each protein family. The functions of the pre-computed database will be introduced in the following sections.

2 Algorithms for combinatorial feature extraction

The algorithms for extracting combinatorial group features are composed of two major processes, one is the interval jumping searching algorithm for extracting fundamental consensus motifs and the other is the center star algorithm for ordered multiple segment alignments. In the interval jumping searching module, each fixed-length segment in a query sequence is transformed into a unique number and allocated in its corresponding interval for matching processes. The main idea of improving
such matching processes is based on the encoded number of a common pattern located in the same interval and the match of an equal digital value representing an exact matched status. In addition to perform the combinatorial feature extraction for all query sequences, the system also provides the exclusive group feature analysis for each subgroup. To approach such function, a hierarchical clustering algorithm is provided in advance based on various combinations of parameter settings. The purpose of clustering operations is to increase correlation among sequences. Using previous consensus motif searching processes, sequences with high similarities can be grouped together to improve alignment results. A statistical model addressing the occurrence rates of consensus patterns is applied to build similarity score matrices. Two sequences with more consensus patterns are more likely to be clustered into a group based on the assumption of a higher order of homologous relationships. The labelled consensus motifs for a grouped subfamily are then aligned according to their indexed positions instead of the original localizations of residues in the primary structure. After the interval jumping searching and clustering processes, one sequence will be defined as the central sequence in each subgroup, which possesses the greatest number of consensus motifs from all other sequences in that subgroup. Based on the labelled consensus segments, the center-star alignment methodology and dynamic programming techniques are applied to each sequence with respect to the central sequence individually. The result of each pairwise alignment contains two indexed sequences with appropriate gaps inserted and is further combined and transformed into a multiple alignment indexing matrix through a multi-way merging process. In the aligning processes, the principle of “once a gap, always a gap” is obeyed throughout the operations (7). Finally, the combinatorial features of each subgroup will be displayed in colour-coded bar representation, and the tools for further verification by Prosite (release 20.4) scanning are also provided in the resulting website. In this research, a hit rate of 73.6% was achieved for frequently occurring signatures when considered the extracted patterns with length greater than 5 residues. In other words, 27.4% identified patterns were not annotated by Prosite yet, and which might serve as undiscovered important motifs.

3 Examples of ordered combinatorial feature database

This is an example of applying proposed database to identify combinatorial group features from protein family of interest. The step-by-step guide covers the basic strategy for efficient searching and analysis. The human RNaseA superfamily are illustrated as an example and the combinatorial features as well as their three-dimensional structures are shown and discussed in the results. In the proposed system, four different approaches are provided to reveal the protein families of interest: entering the name of interested protein family, partial sequence contents, contents of known motifs, or related PDB IDs (if existed). For example, users can type "RNase" in the textbox of "Search by family name", "1GQV" in the textbox of
"Search by PDB IDs", “TLIMQ” in the textbox of “Search by Motifs”, or "RAQWFAIQHISLNPPRCTIAMRAINNYWRCKNQNTFLRTTFANVNVNCGN QSIRCIPHINTLNNCHRSLRFRVPLLH” in the textbox of "Search by sequence contents" in the searching interface. Either way will provide users the related protein families in a tabular format. Figure 1(a) shows that a user copies and pastes partial sequence contents and clicks the "Search by sequence contents" to retrieve the related protein families. Once all the possible protein families are listed in the resulting table, users can select one appropriate protein family to reveal its pre-extracted combinatorial features in various conditions. The existing results can be obtained by clicking on the field of "Combinatorial Features" directly. For example, if a user clicks on "5_2_100" of the entry of RNaseA PF00074 protein family, the system will pop out a window containing its combinatorial features. The symbol of 5_2_100 represents parameter settings for the stored results which represent the fundamental segment length of 5, two-residue tolerance condition, and 100% occurrence rates for whole sequence set are defined for performing combinatorial feature extraction, and the results show that only two sequential segments are found in the “RNase A like” protein family as shown in Figure 1(b). The hierarchical clustered results of this protein family is also prerun and shown in Figure 1(c) and the on-line search interface is shown in Figure 1(d) for customized parameter settings. If the structures of this selected protein family are resolved, then the field of “Structures” will appear a ball-like logo in its corresponding box. Users can click on the logo to display all protein structures. After selecting the structures of interest, users can perform the combinatorial feature analysis and structural alignment based on the extracted combinatorial features through an unpublished constrained multiple structure feature alignment system. The before and after figures for constrained structural alignment of this example are shown in Figure 1(e) and (f) respectively.
Fig. 1. An example of RNase A like protein family for demonstration. (a) the query interface of the proposed database, (b) combinatorial feature of RNase A like protein family, (c) hierarchical clustering tree structure of RNase A like protein family based on consensus motifs, (d) on-line search interface, (e) related protein structures of RNase A like protein family, (f) partial selected protein structures and aligned results.

Acknowledgments. This work is supported by Center for Marine Bioscience and Biotechnology (CMBB), National Taiwan Ocean University and National Science Council (grants NSC 95-2221-E-019 -032 to Tun-Wen Pai, NSC95-2627-B007-003 and NSC95-2622-B-007-001-CC3 to Margaret Dah-Tsyrg Chang) in Taiwan, R. O. C.

References

Positional Clustering of Neighboring Genes in the Zebrafish Genome

Wei Wu1,2, Jirong Peng1, Louxin Zhang2

1 Institute of Molecular and Cell Biology, 61 Biopolis Drive (Proteos), Singapore 138673
2 Department of Mathematics, National University of Singapore,
2 Science Drive 2, Singapore 117543
E-mails: {wwu, pengjr}@imcb.a-star.edu.sg, matzlx@nus.edu.sg

Abstract. Microarray experiments provide deep insight into molecular evolution and how structure and function interrelate in a genome. Our study investigates whether neighboring and clustering genes in the zebrafish genome are co-expressed using the Affymetrix microarray data. We study the positional clustering of genes in the zebrafish genome. A significant trend for large clusters is recognized. The correlation of positional clustering of genes and the co-expression level of neighboring genes is also studied. A positive correlation between the significance of positional clustering and the degree of neighboring gene co-expression is found in the genome.

Keywords: Microarray, Positional Clustering, Zebrafish Genome

1. Introduction

Genes are not randomly distributed on chromosomes. The clustering of co-expressed homologous genes could be explained by the evolutionary history of the genomic region. Spellman et al. [1] found that when genes were organized according to their positions along the chromosome. They obtained Drosophila gene-expression profile using oligonucleotide microarrays. They performed a genome-wide analysis of chromosomal distribution of co-expressed tissue-specific genes in Drosophila. Lee et al. [2] found genes that were reliably co-expressed in multiple human data sets established a high-confidence gene network that was connected by co-expression links, and showed how cluster analysis of the network can reveal functionally coherent groups of genes.

In most of the literatures, testing for non-random clustering of specific genes is done by simulation, which generates the null distribution of the test function. The real value of the test function is then compared with the null distribution. Besides the various simulations, alternative analytic test methods are used by researchers for non-random clustering of specific genes. Neighborhood model is one of these methods, which was studied by Durand and Sankoff [3]. Under their model, two testis-specific genes are defined in a cluster if and only if there is a series of the testis-specific genes locating between them, such that the distance between any two successive testis-specific genes in the series is less than a specified threshold. In an independent study

2. Method

In the zebrafish genome, we perform a genome-wide analysis of the genes, in order to identify the positional clustering of the co-expressing genes. We evaluate the positional clustering of genes using neighborhood model. We define clusters using the distance between two neighboring genes along the chromosome. Two genes are in a cluster if there is a series of genes between them such that the distance between the two neighboring genes in the series is less than a specified distance \( D \). To incorporate the variance of gene density in different regions on a chromosome, we divide a chromosome into segments of length \( L \) so that genes on such a segment are roughly uniformly distributed, and analyze the significance of a cluster within the segment according to the number \( N \) of all the genes in the segment and the values of \( L \) and \( D \).

Assuming the start position of a gene is uniformly distributed in a segment of length \( L \), it falls in an interval \((x, x+D)\) in the segment with the probability \( \frac{D}{L} \). Thus, the number of genes that fall in this interval has a binomial distribution with mean \( N \frac{D}{L} \). In our case, \( D/L \) is smaller than 0.1. This distribution is approximately Poisson with mean \( N \frac{D}{L} \). The Poisson approximation implies that the number of the genes that falls in a randomly chosen interval of length \( D \) has a Poisson distribution with mean \( m = ND/L \), and so the probability that at least one gene falls in this interval is \( 1 - e^{-m} \). Suppose \( g_1, g_2, ..., g_n \) form a cluster in the segment. Then, all the distances of two successive genes are less than \( D \). Thus, the number of genes in a cluster minus one has a geometric distribution with \( p = 1 - e^{-m} \). This implies that the probability that a cluster has \( n \) genes is \( (1-p)p^n \), \( n = 0, 1, 2, ... \). Hence, the \( p \)-value of a cluster with \( n \) genes is \( p^n \). The resulting \( p \)-value indicates the significance of the cluster obtained.

3. Results and Discussion

We conduct the statistical analysis with four different values of \( D \) (25K, 50K, 75K, and 100K), and different \( p \)-values. To study the correlation of mean R values and region sizes, we take the average R values of clustered neighboring genes across all the 25 chromosomes for each region size. After we calculate the average R values of genes across the 25 chromosomes, we get the mean R values for each region size as shown in Table 1. The mean R value for region size 25K is the largest, which is 0.08796. When the region size gets larger, the mean R value decreases gradually. The mean R values of neighboring genes within region sizes of 50K, 75K and 100K are close to each other. It reflects that the neighboring genes within region size of 25K show relatively high level of co-expression, while the genes within region size of 50K or larger do not show the explicit difference of neighboring gene co-expression, as shown in Figure 1.

After the analysis of region size effect to the co-expression level of neighboring genes, we further investigate the correlation of cluster size and the mean R value. Here we calculate the average R values of neighboring genes with different cluster sizes across the whole chromosomes. The result, in Table 2, shows that the mean R
value gets larger when we increase the respective cluster size. Within the specific region size, a larger cluster size means a higher intensity of neighboring genes in the chromosome. Therefore, the result implies a positive correlation between the gene intensity and co-expression level of neighboring genes, for each specific region size.

Table 1: Change of mean R value of neighboring genes across the whole chromosomes, with different region sizes of 25K, 50K, 75K and 100K.

<table>
<thead>
<tr>
<th>Chromosome Number</th>
<th>25K</th>
<th>50K</th>
<th>75K</th>
<th>100K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1608</td>
<td>0.1211</td>
<td>0.0887</td>
<td>0.0963</td>
</tr>
<tr>
<td>2</td>
<td>0.1116</td>
<td>0.0989</td>
<td>0.0717</td>
<td>0.0648</td>
</tr>
<tr>
<td>3</td>
<td>0.1940</td>
<td>0.0726</td>
<td>0.0589</td>
<td>0.0584</td>
</tr>
<tr>
<td>4</td>
<td>0.0390</td>
<td>0.0780</td>
<td>0.0603</td>
<td>0.0391</td>
</tr>
<tr>
<td>5</td>
<td>0.0948</td>
<td>0.0769</td>
<td>0.0437</td>
<td>0.0517</td>
</tr>
<tr>
<td>6</td>
<td>0.0529</td>
<td>0.0035</td>
<td>0.0091</td>
<td>0.0619</td>
</tr>
<tr>
<td>7</td>
<td>0.1788</td>
<td>0.1127</td>
<td>0.0936</td>
<td>0.0890</td>
</tr>
<tr>
<td>8</td>
<td>0.0242</td>
<td>0.0024</td>
<td>-0.0093</td>
<td>0.0015</td>
</tr>
<tr>
<td>9</td>
<td>0.1064</td>
<td>0.0683</td>
<td>0.0357</td>
<td>0.0172</td>
</tr>
<tr>
<td>10</td>
<td>0.0169</td>
<td>0.0280</td>
<td>0.0230</td>
<td>0.0372</td>
</tr>
<tr>
<td>11</td>
<td>0.0186</td>
<td>0.0471</td>
<td>0.0628</td>
<td>0.0393</td>
</tr>
<tr>
<td>12</td>
<td>0.2051</td>
<td>0.1446</td>
<td>0.1361</td>
<td>0.1633</td>
</tr>
<tr>
<td>13</td>
<td>0.0515</td>
<td>0.0483</td>
<td>0.0500</td>
<td>0.0476</td>
</tr>
<tr>
<td>14</td>
<td>0.0687</td>
<td>0.1097</td>
<td>0.0840</td>
<td>0.0455</td>
</tr>
<tr>
<td>15</td>
<td>0.1250</td>
<td>0.0475</td>
<td>0.0533</td>
<td>0.0177</td>
</tr>
<tr>
<td>16</td>
<td>0.0809</td>
<td>0.0341</td>
<td>0.0709</td>
<td>0.0586</td>
</tr>
<tr>
<td>17</td>
<td>0.0563</td>
<td>0.0413</td>
<td>0.0616</td>
<td>0.0641</td>
</tr>
<tr>
<td>18</td>
<td>-0.0635</td>
<td>-0.0282</td>
<td>0.0161</td>
<td>0.0365</td>
</tr>
<tr>
<td>19</td>
<td>0.0429</td>
<td>0.0589</td>
<td>0.0719</td>
<td>0.0314</td>
</tr>
<tr>
<td>20</td>
<td>0.0718</td>
<td>0.0444</td>
<td>0.0450</td>
<td>0.0711</td>
</tr>
<tr>
<td>21</td>
<td>0.1815</td>
<td>0.0320</td>
<td>0.0260</td>
<td>0.0420</td>
</tr>
<tr>
<td>22</td>
<td>0.0484</td>
<td>0.1093</td>
<td>0.0957</td>
<td>0.0837</td>
</tr>
<tr>
<td>23</td>
<td>0.1173</td>
<td>0.0617</td>
<td>0.0756</td>
<td>0.0276</td>
</tr>
<tr>
<td>24</td>
<td>0.0583</td>
<td>0.0727</td>
<td>0.1021</td>
<td>0.1773</td>
</tr>
<tr>
<td>25</td>
<td>0.1232</td>
<td>0.0619</td>
<td>0.0777</td>
<td>0.0873</td>
</tr>
<tr>
<td>Average</td>
<td>0.0879</td>
<td>0.0643</td>
<td>0.0600</td>
<td>0.0583</td>
</tr>
</tbody>
</table>

Table 2: The mean R values of neighboring genes from positional clusters of different sizes across the whole chromosomes, with different region sizes.

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>25K</th>
<th>50K</th>
<th>75K</th>
<th>100K</th>
<th>Raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0644</td>
<td>0.0493</td>
<td>0.0446</td>
<td>0.0366</td>
<td>0.0531</td>
</tr>
<tr>
<td>4</td>
<td>0.0789</td>
<td>0.0348</td>
<td>0.0380</td>
<td>0.0472</td>
<td>0.0530</td>
</tr>
<tr>
<td>5</td>
<td>0.1490</td>
<td>0.0787</td>
<td>0.0712</td>
<td>0.0539</td>
<td>0.0474</td>
</tr>
<tr>
<td>6</td>
<td>0.0461</td>
<td>0.0503</td>
<td>0.0567</td>
<td>0.0464</td>
<td>0.0451</td>
</tr>
<tr>
<td>7</td>
<td>0.2286</td>
<td>0.0362</td>
<td>0.0341</td>
<td>0.0642</td>
<td>0.0428</td>
</tr>
<tr>
<td>8</td>
<td>0.1493</td>
<td>0.1399</td>
<td>0.0437</td>
<td>0.0386</td>
<td>0.0431</td>
</tr>
<tr>
<td>9</td>
<td>0.0929</td>
<td>0.0263</td>
<td>0.0952</td>
<td>0.0351</td>
<td>0.0381</td>
</tr>
<tr>
<td>10</td>
<td>0.5056</td>
<td>0.1033</td>
<td>0.0222</td>
<td>0.0420</td>
<td>0.0393</td>
</tr>
</tbody>
</table>
Our results retrieve the positional clustering of genes in the zebrafish genome. A significant trend for large clusters is recognized, and positional clustering of co-expressed, non-homologous genes indicates the existence of some higher order in regulation of gene expression. The correlation of positional clustering of genes and the co-expression level of neighboring genes is also studied.

References


Co-evolution Analysis of Protein Complexes and Its Applications in Pairing Preferences Prediction

Fangfang Pan†, Dongsheng Che‡, Michelle Momany, Liming Cai, and Russell L. Malmberg*

1Department of Plant Biology, University of Georgia, 30602, Athens, GA, USA
2Department of Computer Sciences, University of Georgia, 30602, Athens, GA, USA
*To whom correspondence should be addressed
† These authors contributed equally to this work

Abstract. Co-evolution between protein interacting partners is a topic of great interest. The dramatic increase in the number of available protein sequences and structures allows the application of mutual information to protein co-evolution analysis. We found 48 hetero interaction pairs in protein complexes with known 3D structures to provide the sequence information for MI analysis. Residues with high mutual information are statistically closer to each other in tertiary structure other than random expectations. This characteristic for co-evolving residues could be used as an additional parameter to help predict protein-protein interaction sites. We analyzed the residue pairing preferences with all the possible co-evolving residues from these 48 complex pairs and generated a matrix to predict residue pairing based upon high mutual information.

1 Introduction

Functionally important residues are often conserved during evolution; these evolutionarily conserved positions are relatively easy to detect computationally. Some important residues can mutate rather frequently during protein evolution if they are accompanied by compensatory mutations happening at other sites (co-evolution). It has been computationally difficult to distinguish these non-conserved but important residues from the background mutations that mutate without co-evolving constraints.

Functional RNA polymers evolve with constraint of keeping stable base pairings of secondary structure. The mutual information statistic has been successfully applied to co-variation studies of ribosomal RNA (Gutell, et al., 1992).

Mutual information (MI) of two discrete random variables x and y is defined as

\[
MI(x,y) = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}
\]

It describes the average information about x contained in y and is expressed as the reduction of entropy given y.

Our goal was to study amino acid residue co-evolution both within and between polypeptides using mutual information. Our approach was to use mutual information to find co-evolving pairs. Among these, we then used known 3D structures to build a prediction matrix of amino acid pairs found in close 3D proximity. A known
weakness of mutual information is its relative insensitivity; however currently there are sufficient 3D structures and related homologous sequences to overcome this limitation in many cases. We also used cross validation to test the sensitivity and specificity of this prediction method.

2 Methods

The overall procedure of our mutual information analysis is shown in figure 1.

![Diagram of the overall approach for MI analysis on multiple sequence alignments]

In previous analyses, positions found with mutual information Z score ≥4 were suggested to be co-evolving residues. We obtained 1,021,410 coevolving pairs, out of which 983,365 pairs had inter-residue distances >8 Å and 37,745 amino acid pairs had inter-residue distances ≤8Å. Amino acid pairing preference matrices were built for these two groups as describe below.

We estimated the contacting likelihood between two amino acid i and j as the log odds of pairing, expressed as

$$\log A = \log \frac{\text{Observed}_{ij}}{\text{Expected}_{ij}}$$

(2)

The score $s$ was calculated for all 210 types of residue pairs comparing the likelihood of these two conditions. $s = \log \frac{A_1}{A_2}$, where $A_1 = \frac{p_i p_j}{p_i \times p_j}$ for the group with distance ≤8Å, and $A_2$ for the group with distance >8Å. Score S is calculated from $S = \sum s_i$.

We used a cross validation approach to test the ability of the matrices to predict physical interactions between co-evolving residues.
3 Results

3.1 Residues with high Mutual Information (MI) tend to be closer than average.

We studied the distance distribution between pairs of residues with high mutual information. In our study the physical distances between two co-evolving residues could vary from minimum 2Å to maximum 255 Å. Although residues that were coevolving could be physically far away rather than close enough to have physical interactions, the average physical distance of co-evolving residues were smaller than the average distance between any two residues. The mean distance of all co-evolving residues was 33Å and the median was 30 Å. The physical distance of all co-evolving residues was significantly shorter than the background using either t-test or Mann-Whitney-Wilcoxon test (p<0.0001). The physical distance between-subunit co-evolving residues was also significantly shorter when compared to the background distance calculated from all between-subunit residues; its mean distance was 47Å and median was 43Å.

3.2 Amino acid composition for residues with high MI

We tested if some amino acids were more likely than others to be found in co-evolving pairs. Amino acid compositions were calculated for all the 566997 residues that had high mutual information. They were compared to the background composition of all the 6896622 residues in 48 complex pairs (Figure 2). The correlation between the residue composition of co-evolving pairs and the background was 0.95. This was similar to the correlations for analysis comparing protein interfaces with whole proteins, which was usually bigger than 0.8 (Glaser, et al., 2001). This showed there were no significant differences between the compositions of co-evolving residues and that of whole proteins. The two residues that had the greatest composition differences among co-evolving residues and the background were glycine and tyrosine, which had around 15% less than and 47% more than the background composition. Compared to the background, amino acid aspartic acid cystine, histidine, and asparagine had relatively high compositions in co-evolving residues.

3.3 Interaction prediction matrix
We constructed an amino acid pairing-preference matrix for protein-protein complexes to study the patterns of which amino acids pairs were most likely to be co-evolving.

The cysteine-cysteine pair has high pairing preference due to the disulfide bond (Glaser, et al., 2001). However in our study, it had very negative pairing.

The protein surface is usually less hydrophobic than the interior, however it is more hydrophobic for protein-protein interfaces (Korn and Burnett, 1991). Among hydrophobic amino acids, tryptophan and cysteine have extremely high pairing scores. Some other high pairing preferences between hydrophobic residues were leucine-threonine, cysteine-threonine, alanine-tryptophan, and threonine-tryptophan.

Hydrophobic with hydrophilic residues were previously found unlikely to associate in protein-protein interface studies (Glaser, et al., 2001). However, in our analysis, we did find that co-evolving residues tended to have hydrophobic-hydrophilic pairs close to each other. For example, lysine-proline, tyrosine-glutamic acid, isoleucine-arginine, phenylalanine-lysine, and phenylalanine-arginine. We also found some co-evolving hydrophilic residues that had more tendencies to pair with each other. For example, arginine-histidine and lysine-glutamic acid.

4 Summary

We used mutual information to identify patterns of co-evolving amino acid residues in multiple sequence alignments based upon protein complexes whose 3D structure had been determined. We observed distinct patterns of amino acid pairing in co-evolving sites which we expressed in the form of amino acid pairing preference matrices. Co-evolving residues tend to be physically closer to each other than average residues, hence the matrices we constructed may have some value in helping to predict when a set of amino acid residues may be physically close in 3D space.

This work was supported by NIH grant 5R01GM072080-02 to LC and RLM, and NSF grant MCB 0211787 to MM.

Reference

Unbiased Validation of Multiple Linear Regression Tagging

Jingwu He, Jun Zhang and Alex Zelikovsky
Department of Computer Science, Georgia State University, Atlanta, GA 30303
{jingwu,jun, alexz}@cs.gsu.edu

Abstract. Recent research found that it is essential to find a small subset of informative SNPs (tag SNPs) that may be used as good representatives of the rest of SNPs. Informative SNPs can be used for compaction of unphased genotype data. Indeed, recent successes in high throughput genotyping technologies (e.g., Affimetrix Map Arrays) drastically increase the length of available SNP sequences and they should be compacted to be feasible for fine genotype analysis. Traditionally, such SNPs are called tags and the selection procedure is referred as tagging. The decision which SNPs should be typed (also referred as tag SNPs) and which should be inferred is based on how well non-typed SNPs can be predicted from typed SNPs.

Meanwhile, Biologists can detect or collect data only from a small part of population due to the reasons of technology and expense. For some data, we still have no way to detect all of SNPs because they are unobserved to us. So, this will bring a problem for scientists that the estimates accuracy of tag SNPs when constructing the complete human haplotype map. Iles [4] and Weale et al. [7] proposed a procedure of ‘dropping SNP’ to investigate unbias performance of tagging method. Following their ideas, this study investigates how the MLR-tagging for statistically coverage performs under unbias study. The experiment results shows MLR-tagging still select small amount of SNPs very well even without observing the entire SNP in the sample.

Keywords: haplotypes, genotypes, SNP, tag selection

1 Tagging Motivation

Recent research found that it is essential to find a small subset of informative SNPs (tag SNPs) that may be used as good representatives of the rest of SNPs. Informative SNPs can be used for compaction of unphased genotype data. Indeed, recent successes in high throughput genotyping technologies (e.g., Affimetrix Map Arrays) drastically increase the length of available SNP sequences and they should be compacted to be feasible for fine genotype analysis. Traditionally, such SNPs are called tags and the selection procedure is referred as tagging. The decision which SNPs should be typed (also referred as tag SNPs) and which should be inferred is based on how well non-typed SNPs can be predicted from typed SNPs.
Informative SNP selection (Tagging) methods have been initially explored in statistical and pattern recognition community as well as following optimization community. In statistics, tags are required to statistically cover individual (non-tagged) SNPs or haplotypes (sets of SNPs), where the quality of statistical covering is usually measured by correlation, e.g., find minimum number of tags such that for any non-tag SNP there exists a highly correlated (squared correlation $R^2 > .8$) tag SNP [1, 2]. In the optimization community, the number of tags is usually minimized subject to upper bounds on prediction error measured as how non-tag SNPs can be predicted from the tag SNPs.

The generic informative SNP selection problem can be formulated as follows:

Given a sample $S$ of a population $P$ of individuals (either haplotypes or genotypes) on $m$ SNPs, find positions of $k$ ($k < m$) tag SNPs such that one can predict (or statistically cover) an entire individual (haplotype or genotype) from its restriction onto the $k$ tag SNPs.

## 2 Tag SNP Selection Based on SNP Statistical Covering

A. Zelikovsky and J. He [6] showed how to separate the tag selection from SNP prediction. Following their work, we first define SNP statistical covering algorithm as follows:

A **SNP statistical covering algorithm** $A_k$ accepts as its input the values of $k$ tags $(t_1, \ldots, t_k)$ of a sample $S$. The output of $A_k$ is $R^2$, that is, $R^2$ is correlation coefficient between the non-tag SNPs and $k$ tags.

We select tags by using SNP statistical covering algorithm as follows: We can check each $k$-tuple of tags and choose the $k$-tuple either maximizing average $R^2$ for all SNPs or number of statistical covered SNPs. This manner of exhaustive search is very expensive in terms of running time. We introduced a greedy manner of selection. It starts with the auxiliary tag $t_0$, finds such tag $t_1$ which would be the best extension of $\{t_0\}$ and continue adding best tags until reaching the set of tags of the given size $k$. This produces hereditary set of tags, i.e., the chosen $k$ tags contain the chosen $k - 1$ tags. This hereditary property may be useful in case if the set of tags can be extended. The runtime of greedy manner is $O(knmT)$, where $T$ is the runtime of the SNP statistical covering algorithm.

## 3 Experimental Results

### 3.1 Illis’s validation procedure: Leave-SNP-out

Meanwhile, Biologists can detect or collect data only from a small part of population due to the reasons of technology and expense. For some data, we still have no way to detect them and they are unobserved to us. So, this will bring a problem for scientists that the estimates accuracy of tag SNPs when constructing the complete human haplotype map. Illis [4] and Weale et al. [7] proposed a procedure of ‘dropping SNP’ to investigate unbiased performance of tagging method. Following their ideas, this study investigates how the MLR-tagging for
statistically coverage performs under unbiased study. The experiment results show
MLR-tagging still select small amount of SNPs very well even without observing
the entire SNP in the sample.

The following datasets are used to measure the quality of our algorithms.

**Chromosome 5q31.** The data set collected by Daly et al. [3] is derived from
the 616 kilobase region of human Chromosome 5q31 that may contain a genetic
variant responsible for Crohn’s disease by genotyping 103 SNPs for 129 trios.

**Three gene regions.** Three regions (ENm013, ENr112, ENr113) from 30 CEPH
family trios obtained from HapMap ENCODE Project. The number of SNPs
genotyped in each region is 361, 412 and 515 respectively.

The performance of the new approach was to test MLR method in ile's man-
nner. We leave column out as a small sample. We selected percentage of 10%, 15%
and 20% of entire dataset as our 'observed' data separately and the remainder
classed as 'unobserved'. Then we randomly generated tag $k = 1, 2, \ldots, 8$ and 10
or 30 were selected from 'observed' to calculate $R^2$ to find the maximal average
$R^2$ in 'observed' region and 'unobserved' region. In such way, the average $R^2$
between tags and nonTag SNPs is maximum. Further, we use the tags selected
from the leave-many-out sample to test how good these tags can statically cover
the entire sample. As result, in Daly data, we use one tag can reach $R^2 27\%$ in its
selected sample data, this one tag can cover the 15% of entire data. T represents
tags, V represents visible SNPs and C represents the entire chromosome. In future,
long haplotypes will be our aim. We will take a large dataset, such as 80K to
test.

The length of the region simulated makes little difference to the accuracy of
the results-what is important is the number of observed SNPs from which the
tSNPs are selected. As the number of observed SNPs increases, so the estimates
become more accurate.

We used four datasets to test our result. We leave-many-out as a small sam-
ple, then we select Tag 1..K, in such way, the average $R^2$ between tags and
nonTag SNPs is maximum. Further, we use the tags selected from the leave-
many-out sample to test how good these tags can statically cover the entire
sample. As the number of observed SNPs increases, so the estimates become
more accurate. When there are more than 10 tag SNPs is observed, the relation-
ship between two SNPs is more closed. In ENr113 dataset, we selected percentage
as 10 get $R^2$ as below Fig1. From testing on different dataset, we found the av-
erage $R^2$ in observed small region is bigger than that in entire region. Using 10
tags with average $R^2 0.75$ in visible sample can cover average $R^2 0.63$ in the
entire sample, 20 tags with average $R^2 0.95$ can reach 0.85 in invisible region,
30 tags is for 0.91.

An example of this for a region of different length with 8 SNPs 'observed' is
shown in figure1. Here the average $R^2$ captured at the unobserved SNPs by the
tSNPs selected in the leave-many-out process is 0.949, the average estimated by
the leave-many-out method is 0.833 while the average estimated by using only
the observed SNPs is 0.950. It can be seen from the line chart of figure 2 that
as the number of unobserved markers increases, so the apparent accuracy of the estimate increases. We provide the result using the other dataset as well.

References


Author Index

Irina Astrovskaya, A1, pp.1
Dumitru Brinza, A1, pp.1
Liming Cai, B10, pp.81
Margaret Dah-Tsyr Chang, B8, pp.72
Dongsheng Che, B10, pp.81
Jian-Ming Chen, B8, pp.72
Qiong Cheng, A2, pp.5
Wei-Yao Chou, B8, pp.72
Tom Conley, A12, pp.43
Venu Dasigi, A8, pp.27
Kelly Domico, B1, pp.45
Klaus H. Ecker, A3, pp.9
Tor Fle, B3, pp.51
Glauber Gongcalves, B6, pp.63
Dazhang Gu, A3, pp.9, A12, pp.43
Jingwu He, B11, pp.85
Robert Harrison, B5, pp.59
Orlando Karam, A8, pp.27
Shreekanth Karvaje, A4, pp.13
Brian Keppler, A12, pp.43
Hyunsoo Kim, A5, pp.17
Younhee Ko, A11, pp.39
Baoli Li, A4, pp.13
Jens Lichtenberg, A6, pp.21
Russell Malmberg, B10, pp.81
Michelle Momany, B10, pp.81
M. Taghi Mostafavi, B4, pp.55
Oyun-Erdene Namsrai, A7, pp.23
Tun-Wen Pai, B8, pp.72
FangFang Pan, B10, pp.81
Haesun Park, A5, pp.17
Jinrong Peng, B9, pp.77
Sailaja Pydimarri, A8, pp.27
Ashwin Ram, A4, pp.13
Bharat Ravisekar, A4, pp.13
Gregory Reck, A9, pp.31, A10, pp.35
S. Rodriguez-Zas, A11, pp.39, B2, pp.47
Allan M. Showalter, A12, pp.43
Rajnish Singh, A8, pp.27
Mudita Singhal, B1, pp.45
Brandon Smith, B2, pp.47
Bruce Southey, A11, pp.39, B2, pp.47
Steinar Thorvaldsen, B3, pp.51
Timothy Tickle, B4, pp.55
Yunfeng Tie, B5, pp.59
Martha Torres, B6, pp.63
Iosif Vaisman, A9, pp.31, A10, pp.35
Cristianno Vieira, B6, pp.63
Andy Ju An Wang, B7, pp.68
Hao Wang, B5, pp.59
Hsin-Wei Wang, B8, pp.72
Lonne Welch, A3, pp.9, A6, pp.21, A12, pp.43
Irene Weber, B5, pp.59
Nils P Willassen, B3, pp.51
Sarah Wyatt, A3, pp.9
Wei Wu, B9, pp.77
Jason Yerardi, A12, pp.43
Alex Zelikovsky, A1, pp.1, A2, pp.5, B11, pp.85
Jun Zhang, B11, pp.85
Louxin Zhang, B9, pp.77