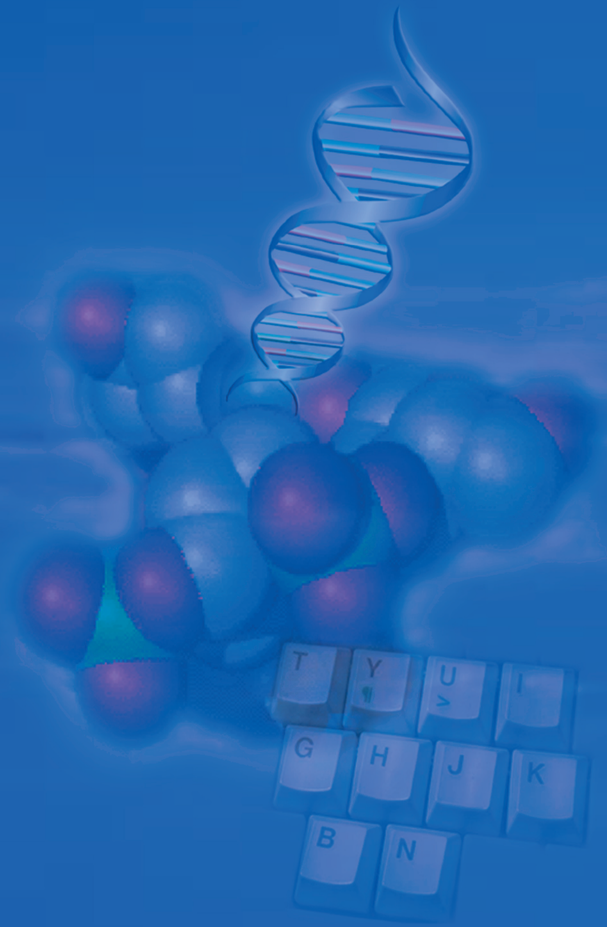
AIBS'02

Annual Israeli Bioinformatics Symposium



14.5.2002

Bioinformatics
Biological
Computing
Unit

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Table of Contents

Symposium Program

Welcome Remarks

Chet, I., President of the Weizmann Institute of Science
Soreq, H., Chair, The National Committee for Scientific Infrastructure

Keynote

SNPs, Protein Structure and Disease
Moult, J., CARB, University of Maryland Biotechnology Institute, USA

Lectures

Inferring Regulation from Gene Expression with Probabilistic Graphical Models
Friedman, N., Hebrew University of Jerusalem

Design Principles of Transcriptional Networks
Alon, U., Weizmann Institute of Science

Sobering Truth about Protein Structure
Trifonov, E., Haifa University

Invading Numts and Resurrected Alus: Evolutionary Dynamics of Junk DNA
Graur, D., Tel Aviv University

A New Look at Cancer Proteomics and Immunology by Large-Scale Analysis of MHC Peptides
Admon, A., Technion

Using the Building Block Folding Model to Reduce the Computational Complexity of Protein Folding
Nussinov, R., Tel Aviv University

Parallel Workshops

Genomics & Medical Bioinformatics - Beckmann, J.S., Chair

Proteomics & Bioinformatics - Edelman. M., Chair, Sobolev,V., Co-chair

Intellectual Property & Bioinformatics - Gressel, J., Chair

Biology Applications for Computational & Mathematical Methods - Domany, E., Chair

Panel Discussion

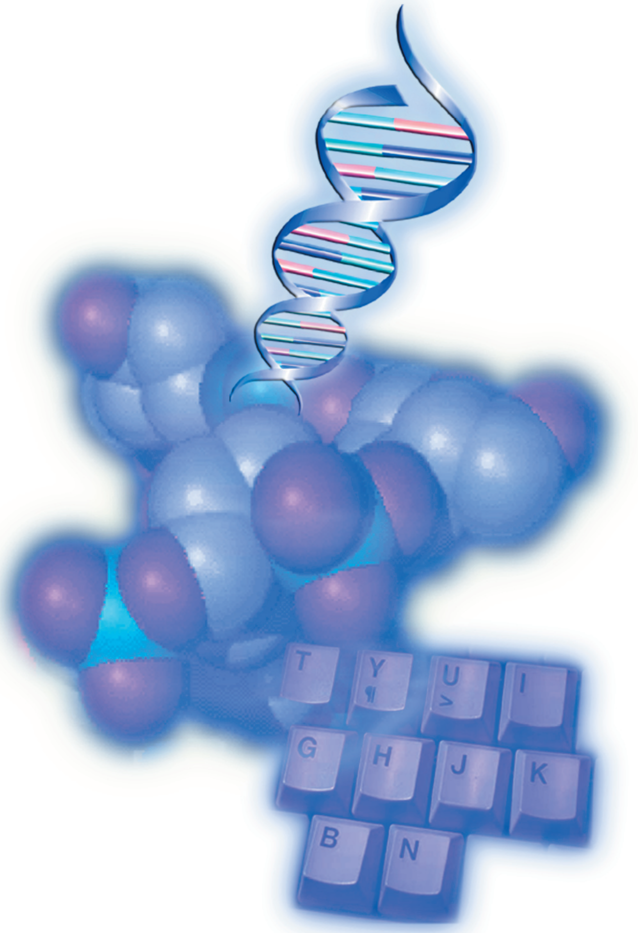
Will Biology become Bioinformatics? - Lancet, D., Chair
Panel: Pilpel, T., Linial, M. and Rubin, E.

List of Poster and Workshop Absracts

POSTER ABSTRACTS

WORKSHOP ABSTRACTS

Poster Abstracts



ConSurf: A Server for the Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information - *Award Winning Poster*

Glaser, F.[1], Pupko. T.[2], Paz, I.[1], Bechor, D.[1], Martz, E.[3] and Ben-Tal, N.[1]
[1] Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University
[2] The Institute of Statistical Mathematics, Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan
[3] Department of Microbiology, University of Massachusetts, Amherst, MA, USA

Mutual interactions between proteins and between proteins and peptides, nucleic acids or ligands play a vital role in every biological process. A detailed understanding of the mechanism of these processes requires the identification of functionally important amino acids that are responsible for these interactions.

It is often difficult to determine the three-dimensional (3D) structure of protein complexes, and sometimes only the structures of the unbound proteins are available. Moreover, crystal contacts between proteins are not always indicative of biologically relevant interactions. Thus, it is common to carry out tedious mutagenesis studies to determine functionally important residues. Because of the amount of work required to determine the functionality of the protein, many entries in the Protein Data Bank have only partial information about their function. The relative fraction of such entries is expected to increase rapidly due to recent high throughput studies to determine protein structures.

Recently, we developed algorithmic tools for the identification of functionally important regions in a protein with known 3D-structure, by estimating the degree of conservation of the amino acid sites within its close sequence homologues. The degree of conservation at each amino acid site is similar to the inverse of the site's rate of evolution; slowly evolving sites are evolutionarily conserved while fast evolving sites are variable. Projecting the conservation grades onto the molecular surface of the protein usually reveals patches of highly conserved (or occasionally highly variable) residues that are often of important biological function.

Here we report the development of a web server, ConSurf, which automates these algorithmic tools and is available to the scientific community at http://bioinfo.tau.ac.il/ConSurf/. Providing a protein structure in PDB format, the server extracts the sequence of the selected polypeptide chain (subunit). It then automatically carries out a PSI-BLAST search for close sequence homologues and multiply aligns them using CLUSTAL W. Alternatively, the user can provide a previously made multiple sequence alignment (MSA). In any event, the server builds a phylogenetic tree consistent with the MSA and calculates the conservation grades taking into account the evolutionary relations between the homologues. The protein, with the conservation grades color-coded onto its surface, can finally be visualized on-line using the Protein Explorer engine.

The ConSurf server enables easy, high throughput studies of proteins with known 3D-structure, and we hope that it will become a standard tool in structural biology studies, in biochemical and molecular biology laboratories.

website: http://bioinfo.tau.ac.il/ConSurf/

Ligand-Protein Docking using Genetic Algorithms and Support Vector Machines

Najmanovich, R., Sobolev, V. and Edelman, M.
Department of Plant Sciences, Weizmann Institute of Science

Goals. Development of a ligand-protein docking algorithm based on genetic algorithms. Prediction of side chain flexibility upon ligand binding using support vector machines.

Background. Ligand-protein docking predictions aim at determining the structure of the ligand-protein complex given the protein atomic coordinates. In certain cases, one may consider both protein and ligand as rigid bodies, in other cases, it might be necessary to consider either the ligand, the protein or both molecules as completely or partially flexible. Genetic algorithms are useful in optimization tasks involving large number of variables and rough landscapes and may be suitable for docking simulations including ligand and side-chain flexibility. A previous work (Najmanovich et. al., 2000) determined that few side-chains undergo conformational changes upon ligand binding. To determine which side chains need to be set flexible, we utilize support vector machines in order to create a classifier system able to predict which side-chains are likely to be flexible during the docking simulation.

Results. We developed a new reproduction technique in our genetic algorithm implementation (Population Boom) that improves convergence. The search algorithm succeeds in finding optimal solutions in both global simulations (searching the whole protein surface) and local simulations (an approximate position for the binding site is known). Our preliminary results using support vector machines to predict side chain flexibility show a classification accuracy of 74.0%±0.9%. Moreover, when used to classify all side-chains on a protein, those side chains predicted to be flexible are present in the surface, in regions of high flexibility such as loops or chain termini.

Clustering of Liquid Chromatography Tandem Mass-Spectrometry Data for Peptide Analysis

Beer, I.[2], Barnea, E.[1], Tamar Ziv, T.[1] and Admon, A.[1]
[1]The Smoler Protein Center, Department of Biology, Technion
[2]IBM Research Laboratory, Haifa, Israel

Liquid chromatography (LC) and tandem mass spectrometry (MS/MS) are commonly combined for analysis and comparison of complex peptide mixtures such as obtained during proteome analysis. The resulting datasets include very large amounts of data combining the full mass spectrum of the peptides and the ms/ms data of selected peptides. A typical mass spectrometer produces hundreds of MS and MS/MS spectra in one run. Even in small-scale proteomics projects, dozens of LC-MS/MS analyses with tens of thousands mass spectra of peptides can be generated, which is beyond the analysis capacity of a human being. The existing peptide identification computer programs only provide a partial solution. We show here how the clustering of similar spectra from multiple LC-MS/MS runs helps manage these data and discover interesting properties of the peptides, the peptide mixtures, and the cells from which the peptides originated. Clustering-based operations contribute to peptide identification by improving spectra quality and providing decision-supporting information. Clustering also facilitates the comparison of peptide mixtures, alleviating the need to identify individual peptides beforehand. In addition, it can be used to correlate the retention time scales of multiple LC runs and to predict peptide retention times from peptide sequences. We implemented the clustering-based methods in a software tool, Pep-Miner. Using the tool, we catalogued the repertoires of MHC Class-I peptides displayed by various human cancer cell types and discovered several cancer-specific peptide candidates for immunotherapy. The methods, however, are not limited to these applications and have the potential to be used for general proteomics.

Sequence Bias in PDB Proteins: Comparison of Dipeptidyl Fragment Counts vs. the Residue Composition

Felder, C.[1], **Einav, U.**[1], **Segal, D.**[1], **Sussman, J.**[1], **Silman, I.**[2], **Beckmann, J.**[3] and **Yakir, B.**[4]
[1] Dept. of Structural Biology, [2] Dept. of Neurobiology, [3] Dept. of Cellular Genetics,
Weizmann Institute of Science ; [4]Dept. of Biological Statistics, Hebrew University of Jerusalem

Examination of sequence frequencies of dipeptidyl units of PDB proteins reveals a bias toward certain sequences relative to what would be expected from a random assembly from the residue composition. A database of the frequency counts of all possible 400 dipeptidyl fragment sequences in PDB proteins was constructed, using the PDB_select list of Hobohm and Sander at 90% homology to eliminate redundant entries. A parallel database of sequence composition was also made. From these data we calculated the observed probability of each dipeptidyl sequence, eg. the raw count divided by the total number of dipeptide fragments in all proteins, against what would be expected from a random combination of residues based on the residue composition. We noted a clear bias in favor of certain dipeptides, such as CH, MM, HP, YW, YC and QQ; and against other dipeptides, such as LW, MW, EC, EP, ES, CV and GP. The ratio of observed over expected probability ranges from 0.7 to 1.5, with an average near 1.0 and std. dev. 1.12. The results suggest that certain combinations of residues may be preferred to facilitate proper folding and function of the proteins.

Rhodopsin Structure is Not an Ideal Template for Modeling G-protein Coupled Receptors: Results from Measuring Variability in Multiple Sequence Alignments

Cherno-Schwartz, S., Rayan, A. and Goldblum, A.
Dept. of Medicinal Chemistry and Natural Products, School of Pharmacy,
Hebrew University of Jerusalem

Conservation relations have been traditionally employed as a basis for constructing 3-dimensional comparative models of proteins. In G-protein coupled receptors (GPCRs), this approach is less obvious due to the lack of experimental structures, except for the single structure of rhodopsin. However, GPCRs are the largest group of drug targets and models for their 3D structures are valuable for drug design.

We have recently developed a novel approach for observing sequence identities in GPCRs, that is based on cumulative conservation in each position and along stretches of residues, starting from either the exoplasmic or endoplasmic helix terminals. The overall average identity of 40 representative GPCRs, 26.3%, is found to be composed of very different values for the endoplasmic (34.5%) and for the exoplasmic parts (17.7%). In five of the 7TM helices (I, II, V, VI, VII), the exoplasmic parts of length 9-17 residues have low cumulative conservation values. We thus expect to find less structural conservation between these helical parts and the corresponding ones of rhodopsin. This should affect any attempt to construct GPCR models based on rhodopsin coordinates. Our recommendation is thus to use the endoplasmic parts of the helical region of rhodopsin as template, and to reconstruct the rest of the structure by different methods, such as de-novo prediction.

New Scoring Function for Modelling Side Chains

Eyal, E., Edelman, M. and **Sobolev, V.**
Department of Plant Sciences, Weizmann Institute of Science

A new scoring function is being developed to predict conformations of amino acid side chains given the backbone conformation. The method is based on a complementarity function (weighted contact surface areas) and a term for intra-residue interactions (derived from the probabilities of rotamers in a rotamer library).

The new function considers solvent interactions in a simple way. So far, it has been tested for individual side chains (with all other side chains held fixed). The results are comparable to those of the latest studies in the field. We are currently working on an algorithm to spatially place several residues simultaneously with the goal of understanding the structure and stability of mutated proteins.

Predicting the (yet) Unknown: The CAPRI Challenge

Eisenstein, M.[1], Ben-Zeev, E.[2], Berchanski, A.[2], Heifetz, A.[2] and Shapira, B.[2]
[1] Department of Chemical Services, Weizmann Institute of Science
[2] Department of Biological Chemistry, Weizmann Institute of Science

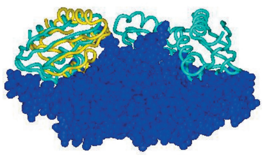
Prediction of the structures of protein-protein complexes is an important and fast developing field in the current structural-genome era. In the past few years we witness a vast increase in the number of publications in this field, which describe a variety of prediction (docking) methods.

All docking methods use the structures of the individual molecules in the complex and try to answer the question "How do they combine to form a complex?" Commonly, the prediction methods are tested first on 'bound structures', attempting to re-assemble protein-protein complexes using the structures of the complexed molecules. Next, the structures of the uncomplexed molecules are used in 'unbound docking', which is a more realistic test of the given docking method. Notably, all the docking methods report much better results for 'bound docking' compared to 'unbound docking'. This is due to a variety of reasons. First, most of the protein-protein docking methods treat the molecules as rigid bodies whereas the structures of the uncomplexed molecules differ from their structure in the complex. In addition, most docking methods include only a part of the interaction energy terms, and often represent them in an approximate form. Nevertheless, success does not elude us and in many cases, good predictions are obtained even in 'unbound docking'. But how do we compare the different docking methods? Each group has its own selection of complexes used for the development and testing of their method; some complexes are easier to predict and others are harder, and each group may or may not publish the results of only some of their prediction attempts.

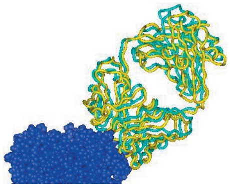
The CAPRI (Critical Assessment of PRediction of Interactions) challenge is a blind docking test that provides a common basis for comparison of different docking methods. The participants are given the structures of the individual molecules and are requested to send their predicted structures of the complexes by a given date, after which the experimental structures are made public. An independent group of assessors tests and compares all the predictions. In the first CAPRI challenge there were 3 prediction targets and 16 predicting groups, which submitted 0-10 solutions per target.

We were the only group that submitted an acceptable prediction for each of the 3 targets.

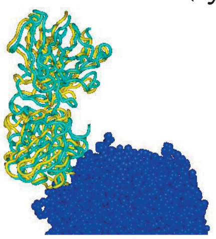
Each figure below presents a superposition of the predicted position of the ligand (yellow ribbon) onto the experimental structure (cyan ribbon).



Target 1



Target 2



Target 3

Role of Hydrogen Bonds and Packing in the Assembly and Stability of Helical Membrane Proteins

Samish, I.[1], Wolfson, H.J.[2] and Scherz, A.[1]
[1] Department of Plant Sciences, Weizmann Institute of Science
[2] School of Computer Science, Tel-Aviv University

The functional assembly and stability of helical membrane proteins is governed by weak inter-helical interactions in a manner still not clearly understood. To study this process, we utilize photosynthetic reaction centers (PRCs) as a model system. PRCs are among the most structurally studied family of membrane proteins enabling *in silico* sequence and structural comparison of this diverse group of hetero-dimer pigment-protein complexes. In vivo support to such analysis is gained by well-established methods for genetic manipulation accompanied by electron-transfer measurements that probe functional stability changes in local environments of the complex.

Multiple structural alignment (MUSTA algorithm) was performed on the different PRC structures: bacterial RC, photosystem I RC and photosystem II RC (PSII-RC). A structural tree was established for the family including a common core to all PRCs – a family lacking sequence identity. Grouping of different amino-acids (AAs) found in this core demonstrated a cluster of 'high-packing' AAs (term defined via the occluded surface algorithm by Eilers *et al*, PNAS 97:5796-5801, 2000), including Gly, Ala, Ser, Thr and Cys found in the 4-helix-bundle center of the complex. Multiple sequence alignment between different RCs confirmed a conserved GxxxG helix-helix packing motif as well as other conserved high-packing motifs in this region. In a search for stabilizing interactions within the packing motifs, inter-subunit hydrogen bond (H-bond) analysis was conducted on each one of the RC structures. A single inter-subunit H-bond was found in the membranous region of each of the RCs.

The role of H-bonds was further studied by combinatorial mutagenesis of the residue donating the putative inter-subunit H-bond in PSII-RC, followed by temperature-dependent biophysical characterization of the mutants (O. Kerner, I. Samish, D. Kaftan, H. Kless & A. Scherz, In preparation). Interestingly, while wild type residue, Ser, and mutants having Cys and Thr in this position showed remarkable stability, incorporation of residues which cannot donate a H-bond, markedly destabilised the complex. Correlation between protein stability (measured by $\Delta G^{\ddagger}_{QaQb}$) and packing values for membrane proteins was high (R^2=0.9) for AAs that contain a polar or charged moiety, and low (R^2=0.06) for AAs that do not contain such a functional group.

Hence, H-bond capacity combined with packing values correlates with membrane protein assembly and stability, enabling deduction of new rules for modeling and design. Future work will include mutagenesis of other AAs involved in electrostatic interactions, molecular dynamics simulation of the mutants at high temperatures, and analysis of additional protein structures.

A Novel Scoring Function for Predicting the Conformation of Pairs of Tightly Packed α-Helices in Transmembrane Proteins

Fleishman, S.J. and Ben-Tal, N.
Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University

Many pairs of helices in transmembrane (TM) proteins are tightly packed. We present a scoring function and a computational methodology for predicting the tertiary fold of a pair of α-helices, such that its chances of being tightly packed are maximized. Since the number of TM protein structures solved to date is small, it seems unlikely that a reliable scoring function derived statistically from the known set of TM protein structures will be available in the near future. We therefore constructed a scoring function based on the qualitative insights gained in the past two decades from the solved structures of TM and soluble proteins. In brief, we reward the formation of contacts between amino acids such as Gly, Cys, and Ser, that are known to promote dimerization of helices, and penalize the burial of large amino acids such as Arg and Trp. As a case study, we show that our method predicts the native structure of the TM homodimer glycophorin A (GpA) to be, in essence, at the global score optimum. In addition, by correlating our results with empirical point mutations on this homodimer, we demonstrate that our method can be a helpful adjunct to mutation analysis. We present a data set of canonical a -helices from the solved structures of TM proteins, and provide a set of programs for analyzing it (http://ashtoret.tau.ac.il/~sarel). From this data set we derived 11 helix pairs, and conducted searches around their native states as a further test of our method. Approximately 73% of our predictions showed a reasonable fit (RMS deviation < 2 Å) with the native structures compared to the success rate of 8%, which is expected by chance.

e-mail: bental@ashtoret.tau.ac.il
website: http://ashtoret.tau.ac.il

Use of the Morphing Graphics Technique to Visualize Conformational Differences Between AChEs from Different Species and Inhibitor-Induced Conformational Changes

Zeev-Ben-Mordehai, T.[1], Silman, I.[2] and Sussman, J.L.[1]
[1] Department of Structural Biology, Weizmann Institute of Science
[2] Department of Neurobiology, Weizmann Institute of Science

There are currently more than 25 AChE structures deposited in the Protein Data Bank, from four different species and/or complexed or conjugated with a repertoire of ligands. A method for sorting and characterizing differences between these structures is presented.

Pairs of AChE structures were aligned using LSQMAN, and rmsd values were calculated. Intermediate models between the two structures were produced by LSQMAN, in Cartesian space, by taking the initial and final coordinates, and interpolating the predicted intermediate coordinates. The intermediate models were then collated into a single QuickTime movie file easily viewable on most computers.

This morphing approach highlighted a conformational difference in loop 319-324 (hAChE numbering) between hAChE and TcAChE earlier reported by Kryger *et al* (*Acta Cryst.* [2000] D56:1385). A similar conformational difference in the same loop between *Dm*AChE and TcAChE was pinpointed utilizing the novel procedure.

A series of movies was compiled, comparing native TcAChE with its complexes and conjugates with a number of inhibitors. These reveal significant inhibitor-induced conformational changes at the top of the active-site gorge. A major conformational change was visualized for the conjugate of *Tc*AChE with diisopropylphosphorofluoridate (DFP).

The simple morphing technique developed thus provides a valuable tool for locating and assessing conformational differences between closely related protein structures.

Structural Genomics of ORFan Genes from *Halobacterium* NRC-1

Shmuely, H.[1], Chehanovsky, N.[1], Dahan, I.[1], Fischer, D.[2], Eichler, J.[1] and Shaanan, B.[1]

[1] Department of Life Sciences, Ben Gurion University

[2] Department of Computer Sciences, Ben Gurion University

With the sequencing of the human genome and the completion of other genome projects, it has become clear that functional knowledge is lacking for proteins encoded by the majority of ORFs. Identifying shared structural motifs offers a strong predictive tool for describing the function of a protein. Unfortunately, our ability to predict the structure of a protein from its sequence alone remains limited. This obstacle is further magnified by the limited number of solved protein structures available for comparative purposes. Structural genomics tries to fill in this gap by solving the structure for as many proteins as possible. In this project we are attempting to solve the structure of ORFs for which no sequence homologs exist (i.e. ORFans) from the halophilic archaeon *Halobacterium salinarium*. In doing so, we hope to reveal known folds in proteins with no sequence similarity or to describe novel protein folds. Accordingly, we have identified 42 ORFans that can be divided into 15 paralogous groups. The structure of each protein was predicted using different bioinformatics web servers. Upon cloning these genes into *Escherichia coli*, encoded proteins have been expressed and purified. After refolding in high salt conditions, the structure of the bacterially-expressed proteins will be determined using X-ray crystallography. In addition, those genes being heterologously expressed in the haloarchaeaon *Haloferax volcanii* will also be crystallized. Based on structural information revealed, we will try to identify the function and biochemical parameters of the proteins.

Genomic ORFans - Past, Present and Future

Siew, N.[1,2] and Fischer, D.[2]
[1] Department of Chemistry, Ben Gurion University
[2] Bioinformatics, Department of Computer Sciences, Ben Gurion University

Sequence ORFans are orphan ORFs (Open Reading Frames) that show no sequence similarity to any other sequence in the databases. ORFans are of particular interest, not only as evolutionary puzzles, but also because little can be learned about them using bioinformatic tools. Thus, the presence of many ORFans does not allow for a full characterization of the genomic content of organisms.

Here we show that the number of ORFans in the first 43 completely sequenced microbial genomes is steadily growing and that after 43 genomes their number is 18,552 (18%) out of a total of 102,114 ORFs. Our data shows that the addition of each new complete genome slowly reduces the number of previous ORFans, but at the same time, the new genome also adds a larger number of new ORFans, and thus the number of ORFans is growing. However, the fraction of ORFans among all ORFs is slowly declining.

Our analysis of size distribution of ORFans and non-ORFans indicates a strong bias towards shorter sequences among ORFans: sequences shorter than 150 residues account for 56% of all ORFans. The fraction of long ORFans in the genomes declines in a rate twice as fast as that of short ORFans. A possible explanation for this bias could be that some short ORFans do not correspond to expressed proteins, or due to limitations of the tools used to identify sequence similarities.

We conclude that the large observed percentage of ORFans reflects a yet unexplained intrinsic property of the genetic material and that further studies aiming at understanding Nature's protein diversity should also include ORFans.

A Novel Amphitropic Motif in the N-terminal Helix of Heterotrimeric G-proteins

Elia,N., Kosloff, M. and Selinger, Z.

Department of Biological Chemistry, The Institute of Life Sciences, Hebrew University of Jerusalem

Heterotrimeric G-proteins relay signals between membrane-bound receptors and downstream effectors. The α subunits of this super-family are anchored to the membrane by one or more lipid modification at their N-termini. These modification can be palmitoylation (also known as S-acylation), myristoylation or both. While the consensus sequence for myristoylation has been well characterized, no sequence determinant for palmitoylation is apparent. We therefore used systematic homology modeling of all different human G_α proteins to look for a three-dimensional structural determinant of palmitoylation rather then a linear sequence motif.

Comparison of the N-termini of this super-family revealed that all α subunits modified only by palmitoylation contain a similar structural motif at their N-terminal helix. This motif is characterized by a prominent positive patch that extends a positive potential well beyond the molecular surface of the protein. Furthermore, this patch is on the opposite side of the N-terminal helix, relative to the face that interacts with the $\beta\gamma$ subunits. Hence, these positive patches are free to interact with the negatively charged inner surface of the plasma membrane. On the other hand, the magnitude of this positive patch is much reduced in α subunits that also undergo myristoylation.

Based on previous results, we suggest that that palmitoylation of G_α proteins requires prior targeting to the plasma membrane. The signal for this membrane localization is therefore either myristoylation or the novel motif that we identified. This signal is further enhanced by interaction of the α subunit with the $\beta\gamma$ complex. The N-terminus of a G_α protein can therefore be described as amphitropic, containing dual signals attracting it to the membrane and enabling it to undergo palmitoylation. As palmitoylation has been shown to modify a plethora of proteins extending beyond G-proteins, this motif could be more widely applicable.

Elucidating the Mechanism of Ras GTPase using Substrate Directed Superimposition

Kosloff M. and Selinger, Z.
Department of Biological Chemistry, The Institute of Life Sciences,
Hebrew University of Jerusalem

G-proteins are a multi-member family of molecular switches involved in a wide variety of essential cellular processes. Ras is a prominent member of this family due to its ubiquitous role in cell proliferation. Despite their functional diversity, all G-proteins behave as conformational sensors of the bound guanine nucleotide. Depending on whether they are charged with GDP or GTP, they change their conformation and consequently their interaction with other proteins in the signaling cascade. G-proteins charged with GTP are in the 'on' state, capable of acting on their downstream effectors. Hydrolysis of the bound GTP (GTPase) switches the G-protein to the 'off' state, characterized by tightly bound GDP. Working out the details of Ras GTPase mechanism is crucial both for understanding the normal function of the protein and for revealing the reasons why and how mutations that interfere with the GTPase reaction are the cause for diseases such as Cholera or Cancer.

Here we use a novel method, Substrate Directed SuperImposition (SDSI), to analyze the abundant structural data available for the active sites of Ras and other G-proteins in different activation states. We argue that this novel approach for comparative analysis of enzymatic mechanisms enables us to compare many structures simultaneously and without bias and to extract new and striking information about their function. Using SDSI and additional data, we propose a new model for the catalytic mechanism of Ras. We suggest that the rate-limiting step in the GTPase reaction is the correct positioning of the conserved glutamine and arginine. This optimal positioning is necessary for these residues to fulfill their role in creating an electrostatic envelope around the substrate, preferentially stabilizing the transition state.

Using engineered Substrate Assisted Catalysis, a unique enzymatic complementation approach, we obtained experimental evidence supporting our new model. We show the data supporting it and discuss the implications for enzymatic catalysis by Ras and other G-proteins and for future therapeutic approaches.

Relevant reference:
"Substrate assisted catalysis - application to G proteins". Kosloff M., Selinger Z. TiBS, March 2001, 26(3): 161-6.

Structural Genomics of ORFan Genes from *Halobacteriun* sp. NRC-1: Homologous Expression of the Cloned Genes in *Halopherax Volcanii*

Dahan, I.[1], Chehanovsky, N.[1], Shmuely, H.[1], Fischer, D.[2], Eichler, J.[1] and Shaanan, B.[1]

[1] Department of Life Sciences, Ben Gurion University

[2] Department of Computer Sciences, Ben Gurion University

With the availabilty of growing number of complete genomes novel approaches for the study of protein function, structure and evolution have begun to develop. For instance, sophisticated computational methods for fold assignment, such as fold recognition or threading, are able to extend the number of assignable sequences. These methods are, however, only useful when the 3D structure of at least one member of a family is known. Therefore there is an urgent need for increasing the number of known folds.

ORFans genes (i.e. ORF's from a given genome that share no sequence similarity with ORF's of other organisms) may represent a cache of novel folds and hence are attractive for structural genomic projects. Accordingly, this project addresses structural description of orfan-encoded proteins from the halophilic archaeon *Halobacterium* sp. NRC-1.

The first stage of the work involved ORFans target selection for structural studies. In such selection, we focused on targets that are most likely to provide a novel functional and/or structural insight. As such, we begin with paralogous ORFan families, as it is reasonable to assume that these indeed correspond to true expressed proteins.

Expression of these genes in another halophilic archaeon is needed for verifying their native structure and in parallel determining their predicted function.

A set of 45 ORFans were selected and divided to 15 groups. Fifteen of these genes were cloned into the shuttle vector pJAM202 under the regulation of an haloarchaeal promoter and tagged by six histidine residues in the C-terminus and introduced into *Haloferax volcanii*. Six of the encoded proteins were expressed and purified on NiNTA resin. Efforts aimed at determining the 3D structure of these proteins, and hence possible functional assignment, are underway.

Rate4Site: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Evolutionary Determinants within their Homologues

Pupko, T.[1], Bell, R.E.[2], Mayrose, I.[2], Glaser, F.[2] and Ben-Tal, N.[2]*
[1] The Institute of Statistical Mathematics, Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan
[2] Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University

Motivation: A number of proteins of known three-dimensional (3D) structure exist, with yet unknown function. In light of the recent progress in structure determination methodology, this number is likely to increase rapidly. A novel method is presented here: “Rate4Site”, which maps the rate of evolution among homologous proteins onto the molecular surface of one of the homologues whose 3D-structure is known. Functionally important regions correspond to surface patches of slowly evolving residues.

Results: Rate4Site estimates the rate of evolution of amino acid sites using the maximum likelihood (ML) principle. The ML estimate of the rates considers the topology and branch lengths of the phylogenetic tree, as well as the underlying stochastic process. To demonstrate its potency, we study the Src SH2 domain. Like previously established methods, Rate4Site detected the SH2 peptide-binding groove. Interestingly, it also detected inter-domain interactions between the SH2 domain and the rest of the Src protein that other methods failed to detect.

Supplementary Information: multiple sequence alignment of homologous domains from the SH2 protein family, the corresponding phylogenetic tree and additional examples are available at http://ashtoret.tau.ac.il/~rebell

* Correspondence should be addressed to:
Nir Ben-Tal, Tel (972-3) 640-6709
Fax (972-3) 640-6834

The authors contributed equally.
e-mail: tal@ism.ac.jp
e-mail: rebell@ashtoret.tau.ac.il
e-mail: fabian@ashtoret.tau.ac.il
* e-mail: bental@ashtoret.tau.ac.il

Rate4Site can be downloaded at:
website: http://ashtoret.tau.ac.il/

bis-Acting Galanthamine Derivatives as Improved Drugs in the Symptomatic Treatment of Alzheimer's Disease

Greenblatt, H.M.[1], Guillou, C.[3], Badet, B.[3], Guénard, D.[3], Thal, C.[3], Silman, I.[2] and Sussman, J.L.[1]
[1]Department of Structural Biology, [2]Department of Neurobiolgy, Weizmann Institute of Science
[3]Institut de Chimie des Substances Naturelles, C.N.R.S, Gif-sur-Yvette, France

The alkaloid galanthamine (GAL), isolated from the *Amaryllidaceae* family of plants, shows strong, reversible anticholinesterase activity. As such it has been tested as a possible alternative to current anticholinesterases such as Aricept©, used in the palliative treatment of Alzheimer's Disease. GAL is already in use in Austria and has been approved for use in the UK and the USA, under the trade name Reminyl®. It interacts with several residues in the active site of acetylcholinesterase (AChE) at the bottom of the "gorge", including Trp84, which binds the quaternary ammonium group of acetylcholine. In an effort to improve the efficacy of this drug, derivatives have been synthesized with the aim of interacting with both the active site and the second cation-binding site at the top of the gorge of AChE, *viz.*, the peripheral binding site. The crystal structures of complexes of three such compounds with *Torpedo californica* AChE have been solved and refined and will be presented.

Crystal Structure of the Tetramerization Domain of Acetylcholinesterase at 2.3A Resolution

Harel, M.,[1], Dvir, H.[1], Bon, S.[2], Liu, W.Q.[3], Garbay, C.[3], Sussman, J.L.[1], Massoulié, J.[2] and Silman, I.[4]
[1] Dept. of Structural Biology, [4] Dept. of Neurobiology, Weizmann Institute of Science
[2] Laboratoire de Neurobiologie, Ecole Normale Supèrieure, Paris, France; [3] Dept. de Pharmacochimie Moleculaire et Structurale, UFR des Sciences Pharmaceutiques et Biologiques, Paris, France

Tetramerization of acetylcholinesterase (AChE) is achieved by the interaction of 2 peptide motifs: a 40-residue ‘tryptophan amphiphilic tetramerization’ (WAT), at the C-terminus of the catalytic subunit, and a 17-residue ‘proline-rich attachment domain’ (PRAD), localized near the N-terminus of the ColQ collagenic tail polypeptide, with 4:1 WAT/PRAD stoichiometry. The two peptides were produced by chemical synthesis. WAT 21Met was replaced by selenomethionine, to permit collection of multiple anomalous dispersion (MAD) diffraction data. The synthetic WAT and PRAD were mixed at a 4:1 ratio, and crystallized. The monoclinic crystals obtained diffracted to 2.3 Å resolution, and MAD data sets were collected at the synchrotron. The structure was solved with the program SOLVE, which produced a traceable electron density map. The structure was refined to an R-factor of 24.6% with the 2 PRADs seen in full and the 8 WATs having disordered C-termini.

The WAT chains assume an α-helical conformation, and are all parallel. The PRAD has a polyproline II conformation and threads its way antiparallel to the WAT chains. Most of the 3 highly conserved Trp residues in each WAT chain are stacked against the 8 Pro residues or 3 Phe residues of the single PRAD. An AChE tetramer structure can be modeled based on the WAT/PRAD complex structure.

Effective Discrimination of Native Protein Structures using Atom-Atom Contacts

McConkey, B.J, Sobolev, V. and Edelman, M.
Department of Plant Sciences, Weizmann Institute of Science

We have developed a method for quantifying inter-atomic contacts within proteins, and have used this measure to conduct a statistical assessment of contacts within known protein structures. The generated scale of atom-atom contact frequencies is useful in the identification of near-native protein structures from within large sets of simulated structures.

The introduced atom-atom contact measure is based on a Voronoi tessellation procedure, which subdivides a protein into cells with a one-to-one correspondence to protein atoms. The contact definition used integrates the solvent accessible surface and atom-atom contacts into a single measure, allowing them to be compared within a statistical framework. The atomˆatom contact measure was used to extract contact preferences from a training set of 648 non-redundant structures, and a contact-based scoring function derived from this data. The accuracy of the scoring function was tested using established protein decoy sets, including decoys from the recent CASP4 competition (http://predictioncenter.llnl.gov/casp4/casp4.html) and those generated by ROSETTA (http://depts.washington.edu/bakerpg). Each decoy set consists of a target native structure plus up to 2000 simulated structures per target. A total of 112 targets and over 48,000 decoy structures were scored. It was found that the scoring function ranked the native protein first within the decoy sets in over 90% of cases, when isolated protein subunits were used as the target structures. If inter-atomic contacts between protein subunits are considered, the accuracy of the method increases to over 97%. This represents a significant improvement in accuracy over currently available methods. Furthermore, the results show that interactions beyond approximately 2 atom diameters (~6.8 Å) are not necessary to determine the fold of a protein for the cases tested, while it is necessary to include interactions between subunits to predict some structures.

Promoter Recognition by Non-Homogeneous VOMT Models

Ben-Gal I.[1], Arviv S.[1], Shmilovici A.[2] and Grosse I.[3]
[1] Department of Industrial Engineering, Tel-Aviv University
[2] Department of Industrial Engineering, Ben-Gurion University
[3] Cold Spring Harbor Laboratory, NY, USA

We suggest a new class of learning models for patterns classification of DNA sequences. The models are based on the context tree that was originally proposed in [7] for data-compression purpose and later modified in [1][2][3].

The suggested models can be described as varying-order Markov Trees (VOMT). Unlike the fixed-order Markov models, the order of various contexts in the VOMT do not have to be equal and, therefore, is not necessarily fixed to account for the maximum dependence found in the data. As a result, the VOMT obtains a great reduction in the number of parameters that need to be estimated, and a smaller probability for over-fitting. Such reduction in the estimation effort is especially important for database of a limited size.

In the present case study we consider the E. coli supervised promoter-recognition (see, [4], [5], [6]). Based on a given dataset that contains 238 E. coli promoters, the VOMT are trained to distinguish between promoters and non-promoters. In particular a non-homogenous VOMT is constructed for the promoters data, such that a context-tree is built for each position in the promoter sequence. **It is shown that the proposed VOMT achieve superior results in comparison to all previously published results.** In particularly, for a cross-validation experiment with a 99.9% true-negatives (TN) value, one obtains a 48.76% level of true-positive (TP) – an improvement of almost 10% to that of the PWM model. Alternatively, for a threshold which is set to zero, one obtains a 92.32% accuracy level, where TP=89.92% and TN=94.73%. Further examples will be given in the talk.

[1] Ben-Gal I., Shmilovici A., Morag G., (2000) Design of Control and Monitoring Rules for State Dependent Processes, *The International Journal for Manufacturing Science and Production*, 3, NOS. 2-4, pp. 85-93.

[2] Ben-Gal I., Shmilovici A., (2001) “Promoters Recognition by Varying-Length Markov Models”, Artificial Intelligence and Heuristic Methods for Bioinformatics, 30 Sept. – 12 Oct., San-Miniato, Italy.

[3] Ben-Gal I., Shmilovici A., Morag G., Singer G (2001) US Provisional Patent Application No. 60/269,344 filed February 20th 2001

[4] Fickett J. W., Hatzigeorgiou A.G., Eukaryotic Promoter Recognition, *Genome Research* 7:861-878,1997.

[5] Holste D., Grosse I., Buldyrev, S. V., Stanley H. E., and Herzel H. Optimization of Protein Coding Measures Using Positional Dependence of Nucleotide Frequencies. J. *of Theoretical Biology*, 206, 525--537 (2000)

[6] Ohler U., Harbeck S., Neimann H., Noth E., Reese M. G., Interpolated Markov Chains for Eukaryotic Promoter Recognition, *Bioinformatics*, 15,5, 362-369.

[7] Rissanen, J., 1983, A Universal Data Compression System, *IEEE Transactions on Information Theory*, 29(5), 656-664.

Minreg: Inferring and Active Regulator Set for Molecular Pathways

Pe'er, D.[1], Tanay, A.[2] and Regev, A.[3,4]

[1] School of Computer Science, Hebrew University of Jerusalem; [2] School of Computer Science, Tel Aviv University; [3] Department of Cell Research and Immunology, Tel Aviv University
[4] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science

Regulatory relations between genes are an important component of molecular pathways. Here, we devise a novel global method that uses a set of gene expression profiles to find a small set of relevant active regulators and identifies the genes that they regulate. Our algorithm is capable of handling a large number of genes in a short time and is robust to a wide range of parameters.

We use our model to automatically provide functional annotation of regulators and identify the logic of regulation. Our global approach characterizes each regulator according to on significant properties of its set of regulatees. This global interpretation provides both statistical robustness and biological generalization.

We apply our method to a combined dataset of Saccharomyces cerevisiae expression profiles, and validate the resulting model of regulation by cross-validation and extensive biological analysis of the selected regulators and their derived annotations.

Correlated Sequence-Signatures as Markers of Protein-Protein Interaction.

Sprinzak, E. and Margalit, H.
Department of Molecular Genetics and Biotechnology, Hadassah Medical School,
Hebrew University of Jerusalem

As protein-protein interaction is intrinsic to most cellular processes, the ability to predict which proteins in the cell interact can aid significantly in identifying the function of newly discovered proteins, and in understanding the molecular networks they participate in. An appealing approach would be to predict the interacting partners by characteristic sequence motifs that typify the proteins that are involved in the interaction. Valuable insight towards this end can be gained by mining databases of experimentally determined interacting proteins. Conventionally, single protein sequences have been clustered into families by distinct sequence signatures. Here we propose a novel approach for clustering different *pairs* of interacting proteins by *combinations* of their sequence signatures. To identify such informative signature combinations, a database of interacting proteins is required, as well as a scheme for characterizing protein sequences by their signatures. In the current study we demonstrate the potential of this approach on a comprehensive database of experimentally determined pairs of interacting proteins in the yeast S. *cerevisiae*. The proteins are characterized by sequence signatures, as defined by the InterPro classification. A statistical analysis is performed on all possible combinations of two sequence signatures, identifying combinations of sequence signatures that are over-represented in the database of pairs of interacting proteins. It is proposed that such correlated sequence signatures can be used as markers for predicting unknown protein-protein interactions in the cell. Such an approach reduces significantly the search in the interaction space, and enables directed experimental interaction screens.

References:

Sprinzak E, Margalit H. (2001), J Mol Biol. 311(4):681-92.

Can a Minimal Gene Set for Cellular Life be Deduced from Comparisons of Completely Sequenced Bacterial Genomes?

Dagan, T.[*], Mintz, S. and Graur, D.
Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University

The ‘minimal genome’ approach attempts to estimate the smallest number of genetic elements sufficient to build a modern-type free-living cellular organism. There are two main approaches for inferring the minimal gene set for cellular life: the experimental and the analytical. The experimentalists use mutagenesis techniques in order to find which genes are essential to sustain cellular life, while the theoreticians infer the minimal-genome-set from known genomes of cellular organisms, i.e., lists of genes that are sufficient to sustain the life of a cell.

In this study, we follow the analytical approach. We estimate the minimum number of genes sufficient to sustain cellular life from the genomes of 30 bacteria species using bioinformatics tools. In our analysis, we make a distinction between free-living bacteria and parasite bacteria since parasitism invariably entails loss of genetic functions in the parasite and a consequent reduction in genome size. We also test the influence of the number of genomes used for the estimation on the final minimal-genome-set size.

Our results show that the minimal-genome-set of free-living bacteria is two times bigger than that of the parasite bacteria. However, the size of the minimal-genome-set is strongly dependant on the number of genomes in the analysis, the more genomes we use the smaller minimal-genome-set we obtain. This notion brings us to the conclusion that the assessment of the analytical approach is strongly biased by the number of genomes used in the analysis and therefore should be treated with caution.

[*] e-mail: tali@kimura.tau.ac.il

A Probabilistic Approach for Inferring Repression Kinetics from Expression Time-Series

Nachman I.[1,2], Ronen, M.[3], Zaslaver, A.[3], Friedman, N.[2] and Alon, U.[3,4]

[1] Center for Neural Computation, [2] School of Computer Science & Engineering; Hebrew University of Jerusalem
[3] Dept. of Molecular Cell Biology, [4] Dept. of Physics of Complex Systems; Weizmann Institute of Science

Many attempts have been done to infer regulatory connections from expression data. A fundamental problem, though, is that the relevant sizes between which such a functional connection exists are the level of activated protein of the regulating gene, and the RNA transcription rate of the regulated gene. In most experiments, the first size is hidden, while the second is measured.

We suggest two methods for learning both the circuitry and the kinetic parameters of a jointly repressed module of genes, given partial knowledge on the circuit wiring. In these methods we explicitly model the dynamics of the hidden repressors. We apply these methods on data from E. Coli promoter activity measurements (Alon, 2001) collected during a recovery period from a stress condition.

The first approach uses a partial probabilistic model, combined with pre-processing of the data. The second uses a fully probabilistic generative model, accounting for different possible noise phenomena.

Both methods use likelihood-based scores, which allow the comparison of different wiring architectures. The models are analyzed in terms of their predictive abilities on test data, and the quality of hidden parameter reconstruction.

From Promoter Sequence to Expression: A Probabilistic Framework - *Award Winning Poster*

Segal, E.[1], Barash, Y.[2], Simon, I.[3], Friedman, N.[2] and Koller, D.[1]

[1] Computer Science Department, Stanford University, Palo Alto, CA, USA

[2] School of Computer Science & Engineering, Hebrew University of Jerusalem

[3] Whitehead Institute, Cambridge, Mass, USA

We present a probabilistic framework that models the process by which transcriptional binding explains the mRNA expression of different genes. Our model unifies the two key components of this process: the prediction of gene regulation events from sequence motifs in the gene's promoter region, and the prediction of mRNA expression from combinations of gene regulation events in different settings. Our approach has several advantages. By learning promoter sequence motifs that are directly predictive of expression data, it can improve the identification of binding site patterns. It is also able to identify combinatorial regulation via interactions of different transcription factors. Finally, the general framework allows us to integrate additional data sources, including data from the recent binding localization assays. We demonstrate our approach on the cell cycle data of Spellman et al.(1998), combined with the binding localization information of Simon et al(2000). We show that the learned model predicts expression from sequence, and that it identifies coherent co-regulated groups with significant transcription factor motifs. It also provides valuable biological insight into the domain via these co-regulated ”modules” and the combinatorial regulation effects that govern their behavior.

Using Structure and Sequence Information for Predicting Transcription Factor Binding Sites

Kaplan, T.[1,2], Friedman, N.[1] and Margalit, H.[2]

[1] School of Computer Science and Engineering, Hebrew University of Jerusalem

[2] Institute of Microbiology, Hebrew University of Jerusalem

In recent years, vast amounts of genomic data are being accumulated at a rapid rate. These data open new avenues for investigation of transcription regulation, and in particular for identifying DNA binding sites of regulatory proteins. The current approaches for predicting transcription factor binding sites are based on multiple alignments of already known sites. Here we propose a structure-based approach for predicting novel binding sites, applicable even to newly discovered proteins.

Our approach begins with solved protein-DNA complexes of some family of transcription factors. Based on the solved complexes, a structural binding model is derived. This model, together with the sequences of DNA-binding proteins and their DNA targets, allows accurate characterization of amino acid-nucleotide interactions. We use an Expectation-Maximization (EM) algorithm to simultaneously determine the exact binding location for each protein-DNA pair, and to estimate the parameters of the amino acid-nucleotide interactions. These parameters can now be used for prediction. Given a transcription factor of that family, the method allows the determination of a specific nucleotide profile of its binding site, even in the absence of previously known targets.

Here the method is applied to the Cys2His2 zinc-finger family. We show the compatibility of our parameters with experimental results, and demonstrate the predictive power of the algorithm.

Tomato Phenomics Web Project

Raskind, A.[1], Gur, A.[2], Sobolev, V.[1], Zamir, D.[2] and Edelman, M.[1]

[1]Dept. of Plant Sciences, Weizmann Institute of Science

[2]Dept. of Field Crops, Vegetables and Genetics, Faculty of Agriculture, Hebrew University of Jerusalem

Phenomics may be described as the integrated informatics study of phenotypic and genotypic information to better understand the complex relationship between the two. This relationship often involves a complex network of gene interactions and quantitative trait loci that considerably hinders the use of classical genetic approaches. Integration of phenotypic and genomic analyses may provide important clues for understanding the flow of information from genome to phenotype. During the last few years, the Israeli Tomato Genome Project has generated a significant amount of data concerning different aspects of tomato biology. The current bioinformatics project aims at introducing the phenomics concept into tomato genomic research. Our goal is a public web-based tool that will facilitate data mining and experimental design by combining genetic, phenotypic, biochemical and physiological data in a single relational database.

Our work concentrates on analysis and presentation of phenotypic data and linking them to the existing genomic resources. This includes web-based tools for statistical analysis of experimental data and mapping of Quantitative Trait Loci. Currently, the system works with a dataset from tomato introgression lines and allows performance of the following operations: comparison of different genotypes for a given quantitative trait; analysis of quantitative trait variation within the whole population; quantitative trait profiling of genotype (comparison of all measured quantitative traits to the parental cultivar); analysis of correlation between the traits.in different years and environments.

Selection of input data and form of output are highly customisable. The user may create a custom sub-list of genotypes to work on. Analysis of results can be viewed in graphic and tabular formats. Publication quality graphics (PDF) and printer-friendly versions of the tables are built into the system. The system is founded on high-quality public-domain software. The combination of an Apache web server, PHP server scripting language and MySQL relational database server provides the means for building a dynamic database-driven interactive web site. It also readily allows inclusion of additional modules as the need develops (discussion forums, message boards, protocol books, etc). Statistical analysis is performed by R-package (www.r-project.org), an extensive and expanding set of analysis tools in combination with elaborate graphic capabilities.

We are currently developing the following capabilities for the system: mapping of QTLs to chromosomes; data submission system; integration of data from mutagenesis experiments; additional options for statistical analysis; repository of relevant methods and protocols.

website: http://bioinfo2.weizmann.ac.il:8080/tomato/PHP/index.php

Detection of Regulatory Circuits by Integration of Protein-Protein and Protein-DNA Interaction Data

Yeger-Lotem, E.[1,2] and Margalit, H.[2]
[1]Department of Computer Science, Technion
[2]Dept. of Molecular Genetics and Biotechnology, Faculty of Medicine, Hebrew University of Jerusalem

The post-genomic era is marked by huge amounts of data generated by large-scale functional genomic and proteomic experiments. These provide various types of genome-scale information, such as binding sites for transcription factors, mRNA expression levels, protein-protein interactions, and protein localization. A major challenge is to integrate these various types of information in order to reveal the intra- and inter-relationships between genes and proteins that constitute a living cell. We present a novel application of classical graph algorithms to integrate genome-wide data on regulatory proteins and their target genes with protein-protein interaction data. We demonstrate how integration of these two types of information enables the discovery of simple as well as complex regulatory circuits that involve both protein-protein and protein-DNA interactions. By applying our approach to data from the yeast Saccharomyces cerevisiae we were able to identify known simple and complex regulatory circuits and to discover many putative circuits. The computational scheme that we propose may be used to integrate additional functional genomic and proteomic data and to reveal other types of relations, in yeast as well as in higher organisms.

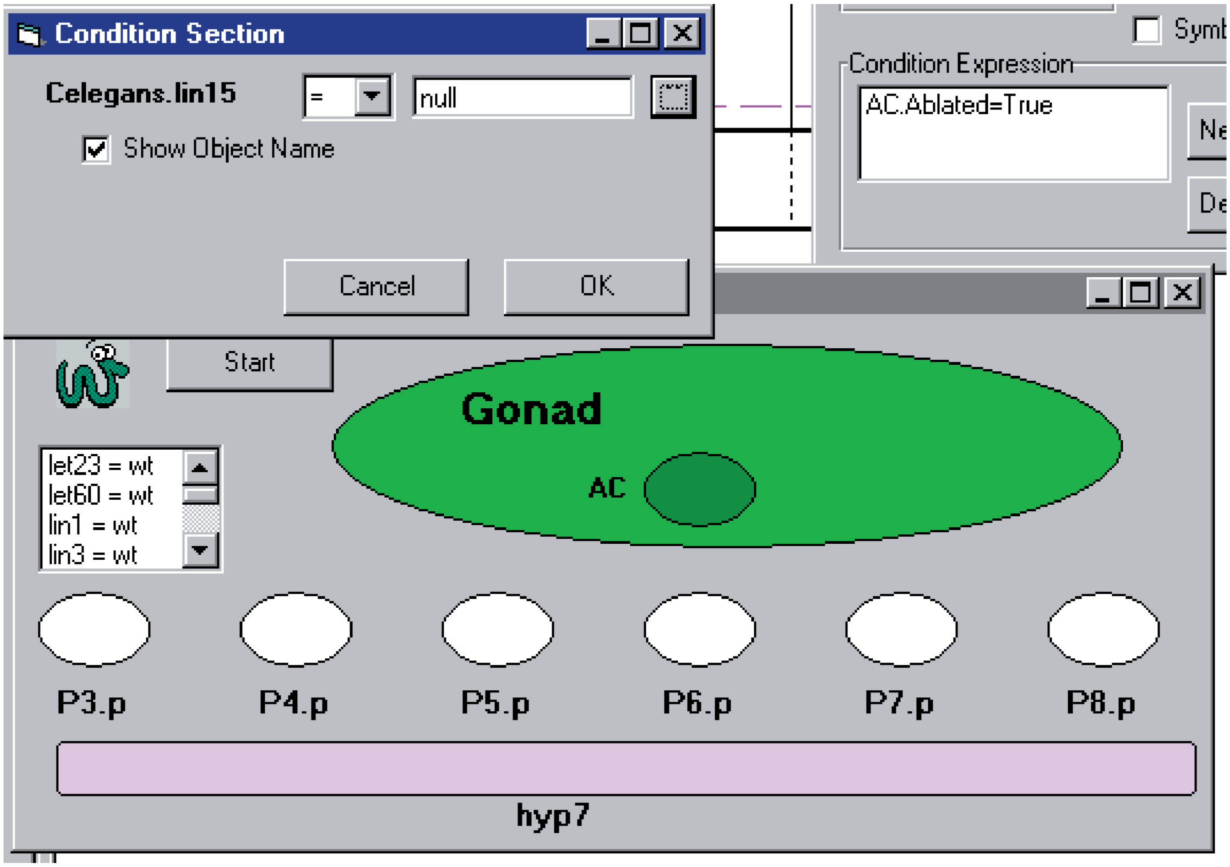
Formal Modeling, Simulation and Analysis of *C. elegans* Development

Kam, N.[1,2], Marelly, R.[1], Kugler, H.[1], Cohen, I.R.[2], Harel, D.[1], Pnueli, A.[1], Stern, M.J.[3]
and Albert Hubbard, E.J.[4]

[1] Dept. of Computer Science and Applied Mathematics,[2] Dept. of Immunology, Weizmann Institute of Science
[3] Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
[4] Department of Biology, New York University, New York, NY, USA

The field of developmental genetics is entering a new phase, in which the synthesis of information from many sources will be necessary to gain a deeper understanding of how various tissues, cells, biochemical interactions and genetic networks collaborate to form a functional organism. The purpose of this project is to model and rigorously simulate and analyze a particular biological system – the *C. elegans* egg-laying system – using languages, methodologies and tools developed by computer scientists for the reliable development of highly reactive computerized systems. The model will incorporate existing anatomical, genetic and biochemical data pertaining to the development and function of (i) the gonad, (ii) the vulva, (iii) the uterine and vulval musculature, and (iv) the hermaphrodite specific neurons (HSNs). We concentrate on an object-oriented approach using the visual language of statecharts for specifying behavior, and tools such as Rhapsody for model execution and analysis. In previous work, we have successfully applied this language and tool to the biological phenomenon of T cell activation. The T cell activation model served as a feasibility test and integrated phenomena associated with cell-cycle control, cell fate, cell behavior and location. We are now in the midst of a far more ambitious effort, involving more complex biological phenomena that will incorporate additional aspects of development, including cell fate acquisition, cell migration, axon guidance, and apoptosis. In principle, our model will eventually handle virtually all aspects of development, ultimately allowing our results to be extended to and used by the entire *C. elegans* community, and will apply to other systems too.

As a first stage, we aim at formalizing the existing genetic, biochemical and anatomical data from the biological literature into a set of live sequence charts (LSCs). These LSCs capture the behavior of the system in terms of inter-object behavior, describing the interaction between objects as scenarios. LSCs enable the user to distinguish between scenarios that can occur in the system, scenarios that must occur in the system, and ones that are forbidden ("anti-scenarios"). Within this aim is the development of a graphical user interface (GUI) for the *C. elegans*. This part of the project will use a recently developed system called the Play-Engine, which enables the user to input the behavioral information in a user-friendly way, and to execute it too. Thus, non-computer scientists can enter biological data in ways in which they are accustomed to representing their system. This will become a critical point regarding one of the future plans of this project: enabling the entire *C elegans* community to 'play-in' experimental data into a behavioral database of LSCs.



Condition Section

Celegans.lin15 = null

☑ Show Object Name

Condition Expression

AC.Ablated=True

let23 = wt
let60 = wt
lin1 = wt
lin3 = wt

website: http://www.wisdom.weizmann.ac.il/~kam/CelegansModel/CelegansModel.htm

Superlink: A New Program for Exact Genetic Linkage Analysis of General Pedigrees

Fishelson, M. and Geiger, D.
Department of Computer Science, Technion

Genetic linkage analysis is a useful tool for mapping disease genes. It allows one to use statistical tools to associate functionality of genes to their location on the chromosome. Generally speaking, this analysis uses a probabilistic model of inheritance of genetic materials and applies it to data in the form of pedigrees, where some of the individuals are annotated with information on the trait of interest and information on their genetic makeup. As highly-informative genetic marker maps have been developed, multipoint linkage analysis has become a crucial part in linkage analysis studies due to its supremacy on pairwise linkage analysis for locating genes and detecting linkage. However, the computational complexity required to perform such calculations increases exponentially due to the large number of markers that participate in the analysis, the high polymorphism of the markers under study, the size of the pedigree, and the number of untyped people in the pedigree. These factors highly constrain the space and time requirements of existing programs. Some programs fail to run as the number of markers, the degree of polymorphism of the markers, or the size of the pedigree increase. Other programs can handle a large number of markers but can only analyze small pedigrees. We have addressed the increasing need for a program that performs multipoint likelihood calculations on general pedigrees with a higher number of polymorphic markers. We implemented our algorithms in a computer program, called Superlink, that computes pedigree likelihood for complex diseases in the presence of multiple polymorphic markers in fully general pedigrees, taking into account a variety of disease models. Superlink compares favorably with current linkage software with regards to the following criteria: functionality, speed, memory requirements and extensibility. This can be seen from the experimental results described below.

Currently, there are two main approaches to computing pedigree likelihood exactly: Elston-Stewart [3] and Lander-Green [5,6,7]. Both of these algorithms are variants of variable elimination methods [2,16] that depend on different strategies to finding an elimination order. The complexity of the Elston-Stewart algorithm is linear in the pedigree size (for pedigrees with a simple structure) but exponential in the number of markers. On the other hand, the Lander-Green method is linear in the number of markers but exponential in the number of individuals. In Superlink, we used the framework of Bayesian networks as the internal representation of linkage analysis problems [4]. Using this representation allows us to give a unified treatment to both approaches and to handle a wide variety of linkage analysis problems. Whenever feasible, we use variable elimination alone to calculate the likelihood of the pedigree. Otherwise, our algorithm combines variable elimination with conditioning (a divide and conquer approach) to achieve the best time-space tradeoff given the memory available for the linkage analysis problem. The crucial point of the algorithm is that conditioning is performed only after some steps of variable elimination, when the memory requirements are about to exceed the limitations. Such conditioning often applies only to parts of the Bayesian network and thus, computations in other, unrelated, parts of the network are not repeated unnecessarily. The elimination order is chosen automatically according to the parameters of the specific linkage problem. For small pedigrees with a large number of markers, the algorithm chooses a peeling order, based on the Lander-Green approach, that proceeds locus after locus. For large pedigrees with a few markers, the algorithm chooses an Elston-Stewart style elimination order which "peels" one nuclear family at a time. Other linkage problems are handled by finding a good elimination order. Often the program chooses an elimination order that is a combination of these two extreme known choices of ordering.

Another crucial feature of our program is the preprocessing step performed on the Bayesian network that reduces the range of values that are feasible for each variable given the data. This step often has a large impact on the memory and time requirements of the calculations. Superlink allows for analysis of sex-linked traits and also allows for a disease phenotype to be under the control of two loci [11, 14, 15].

We have run several experiments to compare our program to some of the leading linkage programs currently, Fastlink [1, 8, 12, 13], Genehunter [5, 6, 7] and Vitesse v1.0 [10]. We have not been able, so far, to try Vitesse v2.0 [9] but we have indications that our program outperforms it on all inputs. The running environment on which all experiments were performed was a Sun OS version 5.7 (sun4u) with 2624 MB RAM. In one of the experiments, we used 12 datasets with a medium sized topology taken from a coronary heart disease study and increasing complexity in terms of the number of loci. The pedigree size exceeds the size that can be handled by Genehunter and only the first few files can be run by Fastlink and Vitesse before the memory requirements become too large. Superlink can run on all the files except for the last one on which the computation will require over a 100 hours in order to complete. It is also important to note that, for the files that could be run by Fastlink and Vitesse, the running times are shorter for Superlink. For example, datasetEA2, required 0.39 seconds by Vitesse and 79.32 seconds by Fastlink and only 0.14 seconds by our program. DatasetEA5 required 84.66 seconds by Vitesse and only 1.19 seconds by Superlink. This dataset cannot run on Fastlink. In another experiment we used a medium-sized looped topology. Vitesse doesn't handle looped pedigrees and therefore failed to run on these files. Fastlink can only run on the first file and its running time is 3933.7 seconds, whereas Superlink takes only 2.56 seconds to run on this file. More experimental results, the full paper, data sets, and executables, are available at bioinfo.cs.technion.ac.il/superlink

References

[1] Cottingham, R.W., Idury, R.M. and Schäffer, A.A. 1993. *Am. J. of Hum. Genet.*, 53:252-263.

[2] Dechter, R. 1998. In J.M.I. (Ed.) *Learning in Graphical Models (pp.*75-104). Kluwer Academic Press.

[3] Elston, R.C. and Stewart, J. 1971. *Hum. Hered.*, 21:523-542.

[4] Friedman, N., Geiger, D. and Lotner, N. 2000. *Proc. Sixteenth Conf. Of UAI.*

[5] Kruglyak, L., Daly, M.J. and Lander, E.S. 1995. *Am. J. of Hum. Genet.*, 56:519-527.

[6] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. 1996. *Am. J. of Hum. Genet.*, 58:1346-1363.

[7] Lander, E.S. and Green, P. 1987. *Proc. Natl. Acad. Sci.*, 84: 2363-2367.

[8] Lathrop, G.M. and Ott J. 1990. *Am. J. of Hum. Genet.*, 47(A188).

[9] O'Connell JR. 2001. *Hum. Hered.*, 51(4):226-40.

[10] O'connell, J.R. and Weeks, D.E. 1995. *Nat. Genet.*, 11: 402-408.

[11] Risch, N. 1990. *Am. J. of Hum. Genet.*, 46:222-228.

[12] Schäffer, A.A. 1996. *Hum. Hered.*, 46:226-235.

[13] Schäffer, A. A., Gupta, S.K., Shriram, K. and Cottingham R.W. 1994. *Hum. Hered.*, 44:225-237.

[14] Schork, N.J., Boehnke, M., Terwilliger, J.D. and Ott, J. 1993. *Am. J. of Hum. Genet.*, 53:1127-1136.

[15] Strauch, K., Fimmers, R., Kurz, T., Deichmann, K.A., Wienker, T.F., and Baur M.P. 2000. *Am. J. of Hum. Genet.*, 66: 1945-1957.

[16] Zhang, N.L. and Poole, D. 1994. In *Proc. of the Tenth Canadian Conference on Artificial Intelligence,* 171-178.

The Degenerate Primer Design Problem

Linhart, C. and Shamir, R.
School of Computer Science, Tel Aviv University

A degenerate primer is a primer sequence, in which some positions contain more than one possible base. Degenerate primers can be used in PCR procedures in order to amplify a variety of similar sequences. We define the Degenerate Primer Design problem (DPD, in short) as a combinatorial optimization problem and prove that various restricted versions of it are NP-complete. We develop an approximation algorithm for one of these variants, and discuss its properties. Finally, we describe an experimental scheme for deciphering the human olfactory subgenome, in which the first step is the design of degenerate primers using a heuristic based on our approximation algorithm.

Given a set of DNA sequences, we wish to design a pair of degenerate primers, so that the primers match and amplify (in the PCR sense) as many of the input sequences as possible. The degeneracy of a degenerate primer is the product of the number of possible bases in each position. In order to reduce the probability of amplifying non-related sequences, we require that each degenerate primer has a degeneracy of no more than some pre-defined constant. We focus on a variant of DPD, called Maximum Cover DPD, in which the input strings and the primer are the same length, and we wish to maximize the number of strings that are matched by the primer. We develop an approximation algorithm for this variant of DPD.

DPD was studied and implemented as part of an experimental scheme for analyzing the composition of a large family of genes with conserved regions. Given a subset of known genes, we design degenerate primer pairs, which are used in PCR procedures to amplify fragments of known, as well as unknown, genes of the same family. The fragments are cloned, spotted on a high-density membrane and oligo-fingerprinted. Another novel algorithm, called CLICK, then clusters the clones into groups of similar genes according to their fingerprints. Finally, representatives from each cluster are sequenced and compared to the existing database. The experimental process was implemented for the human olfactory subgenome, which contains some 1% of all human genes. This project is a joint work with T. Fuchs, M. Khen, G. Glusman, T. Pilpel, D. Lancet (Weizmann), B. Malecova, U. Radelof, J. O'Brien, R. Herwig, H. Lehrah (Max-Planck, Berlin), R. Sharan, and D. Shmulevich (Tel-Aviv).

We have implemented a 3-phase DPD algorithm for designing good primers for the amplification of a family of genes. In the first step, we extract non-degenerate primer candidates from the input DNA sequences using an entropy score. In the second phase, we design primers using an efficient heuristic, which is based on our approximation algorithm. Finally, a greedy hill-climbing function improves the primers we obtain. Our work on the olfaction genes indicates that the DPD algorithm is an efficient heuristic that produces satisfactory results for biological data.

Computational Analysis of a New Methyltransferase Family and of the Destruction Box Motif

Reichmann, D.[1] and Pietrokovski, S.[2]

[1] Department of Biological Chemistry, Weizmann Institute of Science

[2] Department of Molecular Genetics, Weizmann Institute of Science

While protein sequences are relatively easy to determine, identification of their function and structure features by experimental methods is an extremely elaborate and complicated task. Here we demonstrate the use of computational sequence analysis methods, in order to identify functional properties of different protein families and function prediction of uknown proteins. This work is based on the identification and comparison of conserved ungapped sequence motifs ("blocks"). The strength of block analysis was tested in two different ways:

1. Identification of conserved motifs in one protein super-family with very weak sequence similarity,

2. Identification of a conserved motif common to different protein families and apparently responsible for one biological process.

Protein family study: *Identification of a new methylatransferase super-family*

Methylases (or methyltransferases) form a large and diverse group of enzymes that catalyze the transfer of a methyl group from S-adenosyl-L methionine (AdoMet or SAM) to a wide range of substrates. Methylases are found in organisms from all three domains of life (eukaryotes, bacteria and archaea), as well as in viruses and phages. X-ray studies show that methylases from all classes have an alpha/beta type fold with common core structure typical to them. Global sequence similarity between methylases from different families is very weak to undetectable. Using block-to-block comparison methods, thorough sequence analysis of known methylase protein families, we identified a new super-family of proteins from bacteria, archaea and eukaryotes, that appears to be a previously unknown class of methylases. Sequence motifs present in all members of the super-family are most similar to the catalytic and SAM-binding sites of DNA methylases. The significance of this study is in two areas. First, the highly ancient origin of the proteins, and their presence in prokaryotes to mammals, suggests an important role. Second, we demonstrate the strength of blocks analysis for suggesting the function of protein families by identifying subtle sequence similarity to proteins with known function.

Protein motif study: *Defining of the Destruction box (D-box) motif*

Proteolysis by ubiquitin and 26S proteasome pathway is a fundamental mechanism for protein turnover, cell cycle control and signal transudation. The anaphase promoting complex (APC), or cyclosome, is a multi-subunit ubiquitin ligase protein, responsible for ubiquitination of cell regulators at the metaphase - anaphase and mitosis - G1 transitions. D-box, destruction se-

quence motif, is essential for APC mediated protein degradation. A sequence pattern for the D-box was previously suggested as RxxLxxx(N/Q). We characterized more accurately the D-box in various known APC subtracts so it can be used for searching for new putative D-box families. D-boxes are typically distinct between different protein families. A D-box block from one family does not always identify the D-boxes from other families at a significant level, making the identification of a new putative D-boxes a difficult task. However, we succeeded in identifying the D-box in the HspII protein that was experimentally verified. We constructed a web server to identify putative D-boxes:
http://bioinfo.weizmann.ac.il/~danag/d-box/main.html
D-box identification by this server is recommended for known APC substrates or proteins already suspected as targets of D-box dependent degradation. Here we describe a computational approach to identify proteins involved in interactions where one protein, or protein complex, binds to many other proteins (one to many). The strategy of this study can be applied to other studies where one-to-many protein interactions are involved e.g., protein phosphorylation, specific protein cleavage, and cell cycle control.

In-silico Analysis of the *Bacillus Anthracis* Virulence Plasmid pXO1

Ariel, N., Zvi, A. and Shafferman, A.

Department of Biochemistry and Molecular Genetics, Israel Institute for Biological Research

Bacillus anthracis is the causative organism of the potentially fatal disease anthrax. Fully virulent forms of B. anthracis carry two large virulence plasmids – pXO1 and pXO2. So far, only few of the putative virulence factors encoded by these plasmids were identified. These include the toxin components Protective antigen (PA), Lethal factor (LF) and Edema factor (EF), encoded by pXO1; and the anti-phagocytic capsule encoded by pXO2.

Bioinformatics analysis of pXO1 plasmid, aimed at identifying potential B. anthracis specific vaccine/drug targets, was carried out. This analysis may also provide better understanding of the involvement of pXO1 in B. anthracis pathogenesis. The previously defined 143 open reading frames (ORFs) sequenced and partially annotated by Okinaka and co-workers (Okinaka et. al., J. Bact. 181: 6509) were subjected to extensive similarity searches (against the nr (non-redundant) and unfinished microbial genome databases, NCBI), motif analysis (PROSITE and E-motif), cellular location (PSORT, SignalP, gram-positive anchoring signal analysis, TMPRED) and domain analysis (CDD, Pfam, Smart). Genes common to related bacilli (B. subtilis, B. halodurans, B. cereus ATCC 14579 and the pBTOXis plasmid of B. thuringiensis israeliensis) were removed. This comprehensive analysis resulted in a significant increase in the number of ORFs with clues as to their function (from the previously reported 34 ORFs to 85 ORFs). Based on the above, a set of 30 ORFs, consisting mostly of secreted or cell-anchored proteins and proteins with function or motifs typical of documented virulence determinants, were targeted for experimental evaluation of their immunogenic potential.

Identification of Novel Small RNA Molecules in the *Escherichia coli* Genome: from *in silico* to *in vivo*

Hershberg R, Argaman L, Vogel J, Bejerano G, Wagner EGH, Altuvia S, Margalit H.
Hebrew University of Jerusalem

Small, untranslated RNA molecules exist in all kingdoms of life. These RNAs carry out diverse functions and many of them are regulators of gene expression. Genes encoding small RNAs (sRNAs) are difficult to detect experimentally or to predict by traditional sequence analysis approaches. Thus, in spite of the importance of these molecules, many of the sRNAs known to date were discovered fortuitously. We developed a computational strategy to search the *Escherichia coli* genome for genes encoding small RNAs. Our method was based on the transcription signals and genomic features, such as location and conservation, that characterize the 10 known sRNAs in *E. coli*. The search was limited to regions of the genome in which no gene existed on either strand. These regions were searched for transcriptional signals (promoter sequences recognized by the major sigma factor of *E. coli* RNA polymerase (σ70), and Rho-independent terminators). Sequences for which the distance between the predicted promoter and terminator was 50-400 bases were compared to genome sequences of other bacteria. Sequences with good conservation were predicted as sRNAs. 23 of the predicted genes were tested experimentally, out of which 17 were shown to be expressed in *E. coli*. The newly discovered sRNAs showed diverse expression patterns and most of them were abundant.

ProLoc: Prediction of 24 Subcellular Protein Locations

Novik, A., Hazkani-Covo, A.* and Levanon, E.
Compugen Ltd., Tel Aviv, Israel
*Current address: Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University

We have developed ***ProLoc***, a program that can accurately predict the sub cellular localization of a protein solely from its amino acids sequence. ProLoc predicts, with high accuracy, the localization of a protein among 24 compartments, the cell organelles themselves, and their membranes. In addition, it divides the membrane proteins into three groups: Type I, Type II, and integral membrane proteins.

To achieve high levels of accuracy, several different approaches were applied concomitantly. Among these were distributions of the protein's length according to compartment, amino acid composition, prediction of trans-membranous regions, recognition of unique patterns that tend to be specific to a certain organelle (such as NLS), signal peptide and anchor modeling and using unique domains from Pfam that are specific to a single compartment.

Testing the program on Swissprot non-redundant, well annotated proteins that were not part of the training set, the sub cellular location of a protein was accurately predicted as the first choice among the 24 compartments in 73% of the cases, and as the second choice in 12% of the instances. When the possibilities are narrowed down to only five compartments (the secretory pathway, transmembrane, nuclear, cytoplasmic and mitochondrial), the predictions are better.

*__Current address:__ Department of Zoology,
George S. Wise Faculty of Life Sciences,
Tel Aviv University, Ramat Aviv 69978, Israel

Sentinels of Disease: A Bioinformatics and Transcriptome Approach to Elucidating Plant Resistance Genes and their Response

Fluhr, R., Davydov, O., Hadrian, O., Kaplan-Levy, R., Rothan, B., Lund, H.H., Sagi, M., Savaldi-Goldstein, S., Ner-Gaon, H. and Zohar, Y.
Department of Plant Sciences, Weizmann Institute of Science

Successful defense against a pathogen requires perception of its whereabouts. In multicellular organisms this ability can be genetic information that is 'hard-wired' into the genome and is called 'innate immunity'. Innate immunity genes are made up of evolutionary conserved motifs. We have discovered new innate immunity genes and combinations by datamining the Arabidopsis genome.

The pathogen transcriptome interacts dynamically with the plant during infection. Microarray screening of normalized libraries has led to discovery of new fungal pathogenicity targets.

Alternative splicing (AS) amplifies genome complexity. We note an increase in AS rate concomitant to the increase in genome size. Does stress-response alter alternative splicing? We have created a cross-species comparative AS database that will be directed to answer the question of AS dynamics.

Function to Sequence Relations of Intein Protein-Splicing Elements

Amitai, G., Winter, E., Belenkiy, O., Caspi, Y. and Pietrokovski, S.
Department of Molecular Genetics, Weizmann Institute of Science

The relation between the function and structure of proteins and their sequence is a basic problem in biology. We approach this by combining computational sequence analysis and experimental research. Our computational studies focus on the analysis and use of conserved protein regions. These studies guide our experimental research of intein protein-splicing elements.

Inteins are selfish genetic elements. They code for proteins that catalyze their excision out of host proteins, ligating the host flanks with a polypeptide bond. This protein splicing activity is autoproteolytic and is not dependant on any host specific factors. Most inteins also include a homing endonuclease domain that mediates the recombination of the intein gene into alleles lacking the intein element. We are interested in intein function and their evolution. We aim at understanding how inteins protein-splice, how they are selected, and their evolutionary origin.

Inteins are very diverse in sequence but all have a protein-splicing activity that is simple to assay. More then 140 inteins are known from bacteria, archaea and lower eukaryotes. A common set of sequence motifs is present in all inteins and crystal structures of three inteins have been determined. Hence, inteins and protein splicing are an excellent system for studying protein sequence/function relation.

Recently we have shown that inteins with highly atypical active site residues can efficiently protein-splice. Specific mutants were created to test our hypotheses on the protein-splicing mechanism of these inteins. We also identified and showed the activity of a unique group of inteins that occur in insect viruses. These are the only inteins known to naturally protein-splice in the cytoplasm of multicellular organisms. We cloned these inteins and showed them to protein-splice in E.coli and in insect cells. One of these inteins has an endonuclease domain and we show it can cut intein-less and intein-containing alleles and thus can probably mediate horizontal genetic transfer of its gene.

Finding Regulatory Elements from Microarray Data

Ben-Zaken Zilberstein, C. and Geiger, D.
Department of Computer Science, Technion

Microarrays yield enormous amounts of data about gene expression patterns in different cells of different species. The mechanism underlying a particular expression pattern is usually not yet known .Therefore, the development of tools that could reveal regulatory elements involved in an anomalous gene expression pattern is important in understanding the mechanisms underlying pathologies such as human cancer.Since our knowledge about the human genome is not yet complete, it is more convenient to develop these tools by investigating a well known system like the yeast genome, as we did herein. Regulatory elements involved in a certain expression pattern could be revealed by comparing the promoters of a co-regulated group of genes in order to find common sequence patterns differentiating this group from the others. This can be done by local multiple alignment of these promoters in order to detect common consensus sequences (Church et al, 2000) or by detection of over-represented oligonucleotide sequences (van Helden et al, 1998).Bothanalysis methods detect new candidate regulatory sites as well as sites that have already been characterized.We developed a computational tool to reliably predict short oligomers as regulatory elements by examining the promoters of a co-regulated group of genes.

Results We used microarray data containing the expression levels of all yeast genes 15 minutes after the introduction of rapamycin into the cell media, which closely simulate starvation of the cell (Hardwick et al ,1999).

We clustered together a group of genes known to be involved in the amino-acid metabolism that were also at least two fold over expressed 15 minutes after the rapamycin was introduced. We hypothesized that the genes in the over-expressed group share common regulatory elements and we searched this group for any such over-represented oligomers. This was accomplished by the comparison of two frequencies for each candidate oligomer. The first was the frequency of this oligomer in the promoters of the over-expressed genes and the second was its frequency in a control group of 1500 randomly chosen promoters .The higher the ratio between these two frequencies, the higher the chance that a certain oligomer has a biological role which may have caused the over-expression. In order to derive a cutoff for the ratio values to help assess their significance , we performed a simulation calculating this ratio for a randomly chosen group of the same size as that of the over-expressed genes. We repeated this simulation 20 times and used the highest ratio value obtained as the cutoff. When analyzing putative over-represented oligomers, only those exhibiting ratios much higher than the cutoff were selected as regulatory elements. We identified the GCN4 binding site with an extremely high frequency ratio of 10.00 whereas the cutoff was 5.00. GCN4 is a master gene involved in amino-acid metabolism and has already been reported as central in the over-expression of amino-acid metabolism genes during cell starvation (Marton MJ, 2001). All yeast genes that are well known to be regulated by GCN4 (e.g. HIS4, TRP4, ILV1, ILV2, ADE4, ARG1, ARG8, HIS3). were also over-expressed as expected. Another regulatory element we isolated was the CPF1 binding site. CPF1 is involved in the regulation of methionine synthesis. All genes that are regulated by this element (e.g. PGK1, CYT1, and MET16) were also over expressed as expected. We performed the same analysis on a group of genes that participate in the carbohydrate metabolism , and identified oligomers that exhibited ratios far above the cutoff (e.g. GGGGACT ,GGGGG, GGGGC and CGCGC). No literature was found regarding those oligomers. We therefore propose them as regulatory elements of the carbohydrate metabolism during cell starvation.

Discussion Our method has succeeded in identifying some important regulatory elements involved in the process of gene over-expression during cell starvation by analysis of data derived from a microarray .This enables us to throw light on the regulatory mechanism underlying the over-expression phenomenon by suggesting some of the elements involved. Our encouraging results are due to two aspects: First, we focus on genes that react quickly to a certain change (e.g. elevation of rapamycin in the cell media) because this suggests they are responding to a relatively specific common cause. This common cause is likely to co-occur with a small over-represented group of regulatory elements suggesting that our method is effective in these cases. Second, we use statistical simulation to derive a cutoff value, reducing the number of false-positives to below 5\%, making our method very reliable.

References

[1] Hardwick J.S., Kuruvilla F.G., Tong J.K., Shamji A.F. and Schreiber S.L.
Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. Proc. Natl. Acad. Sci USA, 1999 Dec 21;96.

[2] Hughes J.D., Estep P.W., Tavazoie S., Church G.M .Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J.Mol.Biol.2000, Mar 10;296(5):1205-14.

[3] Natarajan K., Meyer M.R., Jackson B.M., Slade D., Roberts C. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeas. Hinnebusch A.G., Marton M.J. J. Mol. Cell Biolology 2001, 21(13):4347-68.

[4] van Helden J., Andre B., Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J.Mol. Biology 1998, 281(5):827-42.

e-mail: chaya@cs.technion.ac.il,
e-mail: dang@cs.technion.ac.il

MotiFinder: Improvement of the Search for Protein Functional Sites using Phylogenetic and Physicochemical Information

Wollman, R. and Ben-Tal, N.
Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University

Functionality assignment to proteins is one of the main goals in molecular biology. The classical way to accomplish this involves expansive and time-consuming mutagenesis studies to determine the residues comprising the functional site(s). Accumulative experimental data have been documented in databases such as PROSITE, which are commonly used to suggest putative functional sites in proteins of unknown function. To this end, stretches of residues comprising a functional (e.g., amidation) site in related proteins were aligned and a signature, corresponding to the functional residues, was derived for the site. Signature derivation is error prone. For example, the signature of a particular functional site reflects the currently documented proteins having this site, and a search using the signature might miss a true functional site even if it is only marginally different from the documented signature.

We describe here a novel method for the identification of signature-like putative functional sites in proteins. The new method, implemented in the MotiFinder program, uses current signature definitions but performs a more permissive search. Each putative signature is assigned a score that reflects its physicochemical similarity to the original signature using an amino acid replacement matrix (1). Another score is assigned to the putative signature based on its evolutionary conservation within homologous proteins (2). These two scores are used to estimate the statistical significance of the putative signature. Currently the program only uses signatures that are defined as PROSITE patterns but its extension to other signatures, e.g., Hidden Markov Model-based signatures, is straightforward.

The method is demonstrated on the Apo-Dethiobiotin Synthetase from *E. coli*; MotiFinder identified a known ATP/GTP binding site (3), which was overlooked by PROSITE, BLOCKS and PRINTS scans.

References

(1) Miyata, T., S. Miyazawa, and T. Yashunaga (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219-236.

(2) Pupko, T., R.E. Bell, I. Mayrose, F. Glaser and N. Ben-Tal (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* (in press).

(3) Yang G, Sandalova T, Lohman K, Lindqvist Y, Rendina AR (1997). Active site mutants of E. coli dethiobiotin synthetase: effect of mutations on enzyme catalytic and structural properties. *Biochemistry* **36**, 4751-4760.

Biases and Complex Patterns in the Residues Flanking Protein N-glycosylation Sites

Rubin, E.[1], Ben-Dor, S.[1] and Sharon, N.[2]

[1] Bioinformatics and Biological Computing Unit, Biological Services, Weizmann Institute of Science

[2] Department of Biological Chemistry, Weizmann Institute of Science

Protein glycosylation, in particular of asparagine residues (N-glycosylation) is the most common and most complex reaction that occurs during protein biosynthesis and often affects markedly their physicochemical and biological properties. It has been estimated that over half of proteins in Nature are glycoproteins (Apweiler et al., 1999). The consensus for N-glycosylation, also known as the sequon, is NXT/S; it is abundant in proteins, but only two thirds are glycosylated. The lack glycosylation of some sequons may be a result, at least in part, of the presence or absence of specific residues at or near the sequon. For example, a proline (Pro) at the X position was reported to be prohibitive for glycosylation. Little is known, however, about the influence of other residues at this position, nor of those flanking the sequon, on the efficiency of N-glycosylation (Shakin-Eshelman, 1996).

We extended traditional approaches of sequence analysis to glycosylation sites in several ways. Using the current version of SWISSPROT, in which 602 well characterized, non-redundant N-glycoproteins have been deposited. The analyzed pattern was extended from the traditional 3-mer sequon NXS/T to a 7-mer sequon M_2M_1NXS/TP_1P_2. Based on experimental information on N-glycosylation of specific aspargines deposited in SWISSPROT, 1186 glycosylated and 717 non-glycosylated 7-mer sequons were analyzed. A supervised learning approach was used to identify complex patterns that separate glycosylated and non-glycosylated sequons.

Analysis of the amino acid distribution at each position of the 7-mer sequon revealed biases in all. Glycosylated sequons showed over-representation of Gly in the X position, and of Leu in the P_1 position and under-representation of Pro in these positions. For non-glycosylated sequons, over-representation of Ser was found in M_2, Asp in M_1, Lys and Pro in X, Tyr in P_1 and Gly in P_2; under-representation of Leu was observed in position P_1.

Supervised learning identified two complex patterns. The data-mining tool WizWhy (WizSoft, Israel) was used to analyze the 7-mer sequons, by describing each position as a separate attribute, and providing the glycosylation state of each sequon as the dependent variable. WizWhy identifies complex "rules" or patterns by first identifying biases in single sites, and merging "rules" that together better explain the dependent attribute.

In glycosylated sequons, several sub-patterns were identified, all matching the consensus D/ESNGTLT. Each sub-pattern matched 2-3 amino acids in positions M_{1-2}, X, and P_{1-2} of the consensus. Scanning SWISSPROT for sequences matching any of the sub-patterns, an abundance of sequons were identified with a strong over-representation for yeast proteins. Interestingly, there are only 3 sequons in SWISSPROT that match the complete consensus.

In non-glycosylated sequons, several sub-patterns were also identified, all converging to the consensus SDNKS/TYG. Each sub-pattern also matched 2-3 amino acids in positions M_{1-2}, X, and P_{1-2} of the consensus, and the perfect consensus was found only twice in SWISSPROT. The sub-patterns of this consensus were also found in abundance in SWISSPROT, but no species biases were observed.

To conclude, patterns were identified in the residues flanking N-glycosylated sequons, both simple biases at single positions, and complex patterns spanning the entire 7-sequon that was analyzed. Our results support some observations made in the past on flanking residues, such as the lack of Pro at position X. However, our results failed to support other suggested preferences, such as the favorable effect of Lys, Arg and Ser on glycosylation, or the inhibitory effect of Trp. We also propose complex patterns that may play a role in the specificity of N-glycosylation.

Gene Expression Analysis Tools

Brodsky, L., Kositsky, M., Leontovich, A., Kalaidzidis, Y., Safro, I., Shtutman, M. and Feinstein, E.

Q.B.I. Enterprises Ltd., Ness-Ziona, Israel

We have developed a software package for microarray analysis, named GEA (for **G**ene **E**xpression **A**nalysis). Its core resides on three basic approaches: nested (hierarchical) clustering algorithm, "main vector" algorithm (correspondence of gene clusters to the matrix of distances between samples), and hierarchical bayesian network of dependencies between genes or gene clusters. The package also contains a proprietary module for quality control of cDNA microarray hybridizations.

Nested clustering provides the understanding of the geometry of distribution of gene expression patterns. It distinguishes the genes smoothly spread within a big area from those ones packed in compact clusters. The basic steps of plain clustering on every level of hierarchy are the following:

- Detection of the best surrounding for every gene according to the probability of its density under hypothesis of the uniformity of the distribution of genes in a wider area.
- Extraction of clusters from surroundings by combination of greedy method and k-means algorithm.

The same two steps are applied to gene-representatives of clusters to produce the next level of hierarchy. Some sub-clusters of an upper level cluster are merged if their separation is not statistically valid.

The aim of "main vector" algorithm is to detect which clusters of genes provide the major input into proximity of probes. Genes as vectors in R^k space (k – number of samples) are transformed into $R^{k(k-1)/2}$ space, where half-matrix of distances between samples is also presented as a vector of this space ("main vector"). The sum of all gene-vectors in this space is the main vector, and thus the clusters of genes making an essential input in the matrix of distances between samples can be distinguished.

This technique can be applied for filtering of genes according to some biological hypothesis (for instance, grouping of samples through the arrangement of the hypothetical matrix of distances between them), and to two way clustering of filtered genes and samples. Unlike Principal Component analysis, "main vector" algorithm can work not only with correlation (covariation) distances between samples, but with Euclidean distance and several other metrics as well.

Finally, hierarchical bayesian network with nodes as vectors of $R^{k(k-1)/2}$ space presents dependencies between nodes, which consists either of genes, or clusters of genes and different types of main vectors arranged for a given population of genes.

Quantum Clustering of Microarray Data in a Truncated SVD Space

Axel, I. and Horn, D.
School of Physics and Astronomy, Tel Aviv University

We describe the application of a novel clustering method to microarray expression data. Its first stage involves compression of dimensions that can be achieved by applying SVD to the gene-sample matrix in microarray problems.

Thus the data (samples or genes) can be represented by vectors in a truncated space of low dimensionality, 4 and 5 in the examples studied here. We find it preferable to project all vectors onto the unit sphere before applying our clustering algorithm, the quantum clustering method. Although the method is not hierarchical, it can be modified to allow hierarchy in terms of its free scale parameter.

We test our method on three data sets and obtain promising results. On cancer cell data we obtain a dendrogram that reflects correct groupings of cells. In an AML/ALL data set we obtain very good clustering of samples into four classes of the data. Finally, in clustering of genes in yeast cell cycle data we obtain four groups in a problem that is estimated to contain five families.

Markovian Domain Signatures: Statistical Segmentation of Protein Sequences - *Award Winning Poster*

Bejerano, G.[1], Seldin, Y.[1], Margalit, H.[2] and Tishby, N.[1]
[1] School of Computer Science and Engineering, Hebrew University of Jerusalem
2 Dept. of Molecular Genetics and Biotechnology, Hadassah Medical School,
Hebrew University of Jerusalem

Characterization of a protein family by its distinct sequence domains is crucial for functional annotation and correct classification of newly discovered proteins. Conventional multiple sequence alignment-based methods, such as hidden Markov modeling (HMM), come to difficulties when faced with heterogeneous groups of proteins. However even many families of proteins sharing a common domain contain instances of several other domains, without any common linear ordering. Ignoring this modularity may lead to poor or even false classification and annotation. An automated method that can analyse a group of proteins into the sequence domains it contains is therefore highly desirable.

We apply a novel method to this problem. The method takes as input an unaligned group of protein sequences. It segments them and clusters the segments into groups sharing the same underlying statistics. A variable memory Markov model (VMM) is built using a prediction suffix tree (PST) data structure for each group of segments. Refinement is achieved by letting the PSTs compete over the segments. A deterministic annealing framework infers the number of underlying PST models while avoiding many inferior solutions. We show that regions of conserved statistics correlate well with protein sequence domains, by matching a unique signature to each domain. This is done in a fully automated manner, and does not require or attempt a multiple alignment. Several representative cases are presented. We identify a protein fusion event, refine an HMM superfamily classification into the underlying families the HMM cannot separate, and detect all 12 instances of a short domain in a group of 396 sequences.

Analysis of a PCNA-Binding Site in Diverse Protein Families

Marcus, S. and Pietrokovski, S.
Department of Molecular Genetics, Weizmann Institute of Science

Proliferating Cell Nuclear Antigen (PCNA) is a major factor in DNA replication, DNA repair and cell cycle control mechanisms. It is found in two of the three major domains of life: Archaea and Eukarya. PCNA is an accessory for DNA-processive proteins, such as DNA polymerases, DNA-repair proteins etc. Several PCNA-binding proteins were shown to bind PCNA in a competitive manner, suggesting a common binding site for these proteins on PCNA, hence a common PCNA-binding motif in these proteins. One of these PCNA-binding proteins is the cell-cycle control protein p21Cip1. A synthetic peptide derived from human p21 Cip1 was co-crystallized with PCNA. This offered an insight to the interaction between PCNA and the proteins that bind to it. The sequence motif of the PCNA-binding site that is shared among p21 Cip1 and other proteins was termed: p21-like PCNA-binding site. A sequence pattern for this motif was offered: Q-x-x-h-x-x-a-a, where 'x' is any amino acid, 'h' is a hydrophobic residue and 'a' is an aromatic residue. This sequence pattern is used for searching for new PCNA-binding proteins. However, a search with this pattern gives a large number of unranked hits that includes many false hits. Thus, identifying true hits is very difficult. Multiple sequence alignments are well documented to be superior to pattern and single-sequence queries in database searches.

We present here a block alignment based method for identifying PCNA-binding proteins. The method has better sensitivity (identifying true hits) and selectivity (avoiding false hits) then pattern and sequence searches. We identified new PCNA-binding sites in several families, including some not even known to bind. Some of the sites appear in a different sequence context then known PCNA-binding sites. Our analysis also allows us to study the convergent evolution of these sites and the effect of various selection pressures on them.

The Cellular Immune System as a Gene-Prediction Resource

Altuvia, Y., Lithwick, G. and Margalit, H.
Department of Molecular Genetics and Biotechnology, Hadassah Medical School,
Hebrew University of Jerusalem

One of the major obstacles to gene prediction remains the ability to demonstrate *bona fide* expression in the cell, at both the mRNA and protein levels. Here we demonstrate how the apparently unrelated field of cellular immunology can aid in gene detection and in the confirmation of otherwise hypothetical genes and proteins. The cellular immune system presents, via major histocompatibility complex (MHC) class I molecules, short peptides that are the degradation products of both foreign and self-proteins expressed in the cell. In uninfected cells, these peptides can be viewed as the remnants of translated gene products, providing leads to their source genes. A database of hundreds of individually sequenced peptides eluted from MHC molecules has been organized and is publicly available. We carried out a comprehensive search comparing these peptides to all accumulated human sequence data. These were in the form of proteins, mRNA, expressed sequence tags (ESTs), and human protein and mRNA predictions. Our findings illustrate how these peptides are informative for the identification of new genes, for hypothetical gene verification, for verifying gene expression at the protein level and for supporting splice junctions.

Bayesian Learning of Haplotype Block Variation

Greenspan, G. and Geiger, D.
Department of Computer Science, Technion

Observable haplotype blocks arise from the interaction between recombination hot-spots, bottleneck effects and genetic drift. The presence of recombination hot-spots in human chromosomes has been demonstrated by several recent high-resolution studies of SNP covariation. They separate between stretches of up to 100,000 base pairs in which almost no recombination takes place, so the SNPs lying between hot-spots act a single multi-site allele or `haplotype block'. A bottleneck occurs when a locally-reproducing population is descended from a small group of individuals, for example due to migration. As the new population grows, it will exhibit far less genetic variation within each block than expected for its size. These small populations also undergo significant genetic drift, in which the variation is decreased further by many generations of random mating.

An accurate statistical model of the haplotype blocks present in a chromosomal region can be used to strengthen the power of genetic association analysis, improve the accuracy of general haplotype resolution and further our understanding of the recombination process itself. Empirical studies of populations descended from a bottleneck, confirmed by our simulations, show that the haplotype blocks in a chromosomal region can be modelled by a dynamic Bayesian network. Each hidden variable corresponds to the ancestral source of a haplotype block, with first-order Markovian transition probabilities reflecting the recombination which has occurred at the hot-spots in the intervening generations. Each SNP observed in an individual chromosome depends upon the ancestral block from which it is descended, under a suitable mutation model.

We have developed a general tool which learns this dynamic Bayesian network model from raw SNP data. The problem differs from classical Markov model training in several ways. The location of hot-spots is not given, requiring a selection between 2^(loci-1) possible network topologies. The values for each hidden state must also be inferred, a difficulty compounded by the presence of failed measurements and the fact that only joint SNP measurements from pairs of chromosomes are often available. Utilizing an ML (maximum likelihood) approach leads to over-fitting, producing a model in which there are no recombination hot-spots and too many ancestral haplotypes. So we adopt the MDL (minimum description length) criterion, which seeks to minimize the number of bits required to represent data D with a model M, given by DL(M)-log2(Pr(D|M)).

Starting with no hot-spots, our search strategy iterates over possible hot-spot insertions (or deletions and nudges in later rounds), trying only those operations which improve our score more than their neighbors, repeating until no further improvement can be found. For a particular assignment of hot-spots, the haplotype blocks for each subject are obtained via a hierarchical EM procedure, which handles both joint unphased and failed measurements. The transition probabilities between the discovered blocks are inferred by EM, then block values are iteratively eliminated to further improve the DL score. Tests on both simulated and real-world data demonstrate our method's ability to recover the haplotype block distribution of a chromosomal region from phased or unphased samples. Our algorithm is guaranteed to converge and takes O(loci ^ 2*samples) time.

Future work will focus on improving the decisions made in the block value elimination stage, to deal with data from older populations in which many mutations have taken place. Avenues being explored include modifying the EM iterative procedure for MDL instead of ML, model-based cluster analysis, and phylogenetic tree pruning. An accurate choice of ancestor blocks will also allow our method to estimate site-specific mutation rates from the data observed.

Molecular Basis of Electrophysiological Diversity of Neocortial Interneurons - *Award Winning Poster*

Toledo-Rodriguez M.[1], Blumenfel B.[1], Wu C.Z.[1], Luo J.Y.[1], Mae S.L.[1], Attali B.[2], Markram H.[1]

[1] Department of Neurobiology, Weizmann Institute of Science

[2] Department of Physiology, Medical School, Tel Aviv University

A major challenge in the post-genomic era is to establish the functions of specific genes and combinations of genes. This effort is fundamental to deriving the genetic basis of the structure and function of the nervous system. The immense computational power that underlies the cognitive and adaptive capabilities of the nervous system arises from interactions of a vast number of neurons with a spectrum of complex and electrophysiological behaviors. While it is believed that these electrophysiological behaviors are generated because neurons express different combinations of ion channel genes, the actual expression profile that underlie the behavior of any specific type of neuron, is not known.

Towards this aim, we have generate a comprehensive single cell cDNA library of morphologically and electrophysiologically fully characterized rat neocortical neurons. We have obtained whole-cell patch-clamp recordings from different classes of interneurons and pyramidal cells and derived a comprehensive breakdown of their electrophysiological properties. After the electrophysiological recordings and loading the neurons with the anatomical dye biocytin, we aspirated the cell´s cytoplasm for subsequent RT-PCR testing for the simultaneous expression of more than 50 genes, including more than 30 channel alpha and beta subunits, 3 calcium binding proteins, 10 neuropeptides and 5 synthesizing enzymes.

The obtained data provided new insights in the expression of neuropeptides, calcium binding proteins and neurotransmitter generating enzymes in neocortical interneurons and pyramidal cells and revealed a variety of (co)-expression patterns of the detected mRNAs encoding for ion channel subunits. Furthermore, data analysis led to the correlation of gene expression patterns and electrophysiological behavior, and even allowed for determining how the expression of specific ion channel subunits influences specific electrophysiological features. The mathematical means for correlating the electrophysiological behavior and genetic expression pattern of a cell provide a formula that allows the prediction of either of these characteristics when given the other.

We will present the results of detail correlations between mRNA profiles and electrophysiological features as well as the algorithm derived from this results that allows to predict the electrophysiological behavior of the neuron based on its gene expression profile.

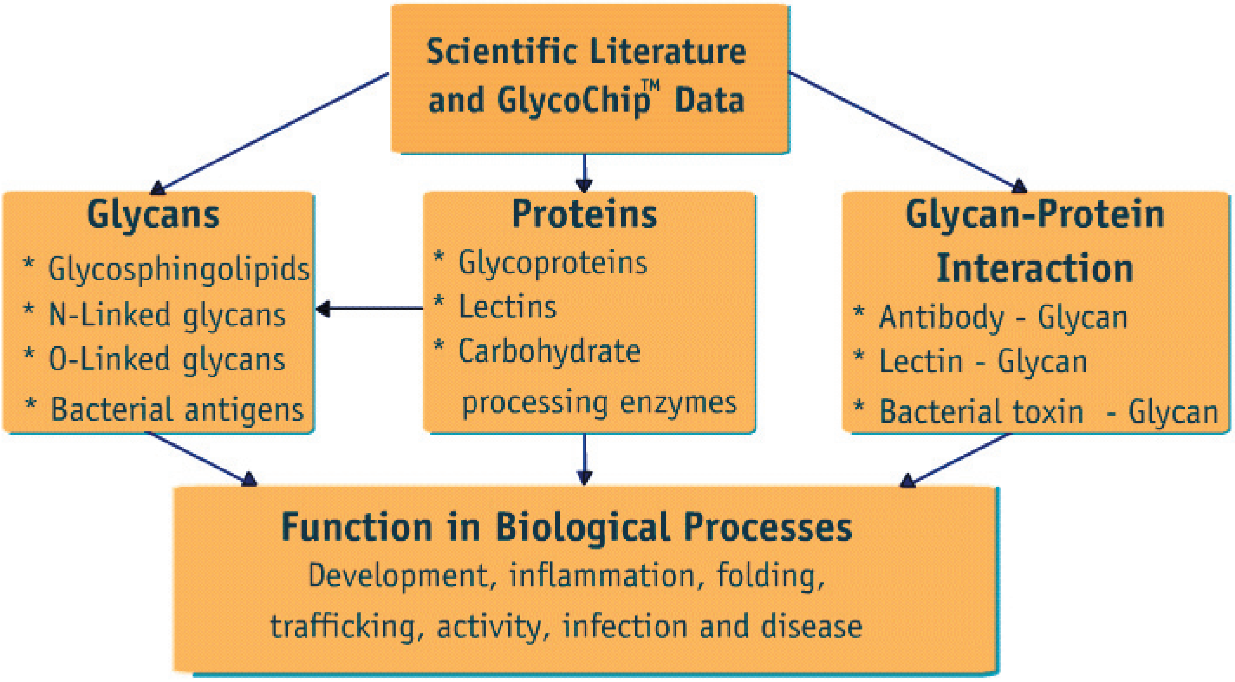
Glycomics Database - A valuable Resource for Glycomics Information

Banin, E., Neuberger, Y., Shulman, M., Gotkine, E., Halevi A., Dukler A. and Dotan N.
Glycominds LTD., Israel

Glycominds has developed the Glycomics Database (www.glycomics.com), which compiles information about glyco-conjugated molecules - including their structures, functions, and interactions with other molecules. This proprietary database contains extensive data on thousands of unique glycans from humans, other mammals, plants and bacteria. The information in the database can be utilized in a wide spectrum of drug discovery and development projects.

All the glycans in the Glycomics database are stored in Linear Code&trade; format, a proprietary nomenclature that enables succinct and systematic description of glycans. Linear Code is extremely comprehensive as it accounts for variability in stereospecificity, ring conformation, modifications, bond conformations, branch patterns, and glycoconjugate molecules. In addition, the Linear Code "language" enables the application of unique technologies for comparative analysis of glycans. Glycominds has developed several glyco-bioinformatics tools for this purpose and their specific utility is dependent on the intended research needs.

e-mail: udib@glycominds.com
website: www.glycomics.com


Scientific Literature and GlycoChip™ Data
Glycans
* Glycosphingolipids
* N-Linked glycans
* O-Linked glycans
* Bacterial antigens
Proteins
* Glycoproteins
* Lectins
* Carbohydrate processing enzymes
Glycan-Protein Interaction
* Antibody - Glycan
* Lectin - Glycan
* Bacterial toxin - Glycan
Function in Biological Processes
Development, inflammation, folding, trafficking, activity, infection and disease

Dialog between Bioinformatics, Molecular and Structural Biology, and Biochemistry Illuminates the Interaction of Functional Domains in a Multi-Enzyme Complex

Bayer, E.A.[1], Barak, Y.[1], Nakar, D.[1], Ding, S.Y.[1], Mechaly, A.[1], Shimon, L.J.W.[2], Frolow, F.[4], Morag, E.[1], Shay, T.[3], Eisenstein, M.[2], Qi, X.[4], Lamed, R.[4], Benhar, I.[4] and Shoham, Y.[5]

[1] Dept. of Biological Chemistry, [2] Dept. of Chemical Services, [3] Dept. of Computer Science and Applied Mathematics; Weizmann Institute of Science, [4] Dept. of Molecular Microbiology and Biotechnology, Tel Aviv University, [5] Dept. of Food Engineering and Biotechnology, Technion

Many cellulolytic microorganisms produce intricate multi-enzyme complexes called cellulosomes that efficiently degrade cellulose - the most abundant organic polymer on Earth. The cellulosomes are composed of a conglomerate of subunits, each of which comprises a set of interacting functional modules. A multi-functional integrating subunit (called scaffoldin) is responsible for organizing the cellulolytic subunits into the multi-enzyme complex. This is accomplished by the interaction of two complementary classes of domain, located on the two separate types of interacting subunits, i.e., a cohesin domain on scaffoldin and a dockerin domain on each enzymatic subunit. The high-affinity cohesin-dockerin interaction defines the cellulosome structure (Fig. 1). The scaffoldin subunit also bears a cellulose-binding domain (CBD) that mediates attachment of the cellulosome to its substrate.

Due to the complexity of this system, progress in this area is dependent on a multidisciplinary approach. Within the past decade, the original contributions of biochemistry to this area have been advanced greatly, first by molecular biology and sequence analysis, and more recently through the benefits of three-dimensional structure determination.

The relevant genes available from sequence databases have been augmented in our lab by sequencing of new cellulosomal genes from other bacteria and by newly emerging genome sequences. Multiple sequence alignment and phylogenetic analysis of the cohesins, dockerins and CBDs have shed light on potentially important residues involved in the functioning of these modules. X-ray and/or NMR structures of recombinant modules provides a structural basis for homology modeling and mapping of the suspected residues. Their involvement in cellulosome function is further examined by site-directed mutagenesis of the desired residues and biochemical analysis of the mutated proteins. Currently, random mutagenesis combined with phage- and cell-display methods

Bayer, E.A., Chanzy, H., Lamed, R. and Shoham, Y. (1998) Curr. Opin. Struct. Biol., 8, 548-557.

Fierobe, H.-P., Mechaly, A., Tardif, C., Belaich, A., Lamed, R., Shoham, Y., Belaich, J.-P. and Bayer, E.A. (2001) J. Biol. Chem., 276, 21257-21261.

Mechaly, A., Fierobe, H.-P., Belaich, A., Belaich, J.-P., Lamed, R., Shoham, Y. and Bayer, E.A. (2001) J. Biol. Chem., 276, 9883-9888.

Mechaly, A., Yaron, S., Lamed, R., Fierobe, H.-P., Belaich, A., Belaich, J.-P., Shoham, Y. and Bayer, E.A. (2000) Proteins, 39, 170-177.

Pages, S., Belaich, A., Belaich, J.-P., Morag, E., Lamed, R., Shoham, Y. and Bayer, E.A. (1997) Proteins, 29, 517-527.

Shimon, L.J.W., Bayer, E.A., Morag, E., Lamed, R., Yaron, S., Shoham, Y. and Frolow, F. (1997) Structure, 5, 381-390.

Shoham, Y., Lamed, R. and Bayer, E.A. (1999) Trends Microbiol. 7, 275-281.

Bioinformatics Support and Training at the Weizmann Institute of Science

Rubin, E., Ben-Dor, S., Chalifa-Caspi, V., Esterman, L., Fichman, S., Frydman, R., Ophir, R., Orr, I., Prilusky, J., Safran, M. and Kahana, C.

Bioinformatics and Biological Computing Unit, Biological Services, Weizmann Institute of Science

To support the ability of Weizmann scientists to use the rapidly growing wealth of biological data and tools, the Bioinformatics and Biological Computing Unit (BBCU) maintains a large professional team. It serves as a knowledge base, by offering consulting services for bioinformatics users and developers. BBCU organizes many bioinformatics training activities such as courses, workshops and seminars in the fields of sequence analysis, genomics, bioinformatics software and database development, and DNA array data processing and analysis. The BBCU also identifies, purchases, and downloads useful data and software, as well as provides information about the availability and usefulness of these tools. The software includes Celera's CDS, Compugen's Gencarta, Accelerys' GCG, and more. Data sources include Genbank, EMBL, Swissprot, NCBI's NR, Transfac and over 50 additional databases. Finally, the BBCU maintains the infrastructure required to facilitate bioinformatics usage, including computers, software, data and mirror sites on a variety of hardware/software platforms. Some of the BBCU services are supported through the Israeli National Node (INN), and are provided to the entire Israeli academic community.

A Data-Mining Approach Applied to Life Sciences

Marcus-Kalish, M.[1], Bonne-Tamir, B.S.[2], Kalid, O.[3] and Freeman, A.[3]
[1] Interdisciplinary Center for Technology Analysis and Forecasting, Tel Aviv University,
[2] Medical School, Tel Aviv University,
[3] Department of Biotechnology, Tel Aviv University

A data-mining tool for analysing data and issuing predictions will be presented.

The motive behind WizWhy, the data-mining algorithm, developed with Abraham Meidan, was the need to reveal, especially in life sciences, the underlying rules behind specific phenomena.

Unlike other available tools for data mining or tools for prediction (such as neural networks, decision trees or genetic algorithms), the aim of this algorithm is to reveal ALL inter-variable relationships in order to construct a theorem and unravel the rules behind the inspected phenomena.

The association rules approach, used in this algorithm, is the only one which is committed to revealing all the if-then rules (that meet pre-defined thresholds), in regard to the rule's confidence level (i.e., probability) and support level (i.e., number of cases).

Further more, the if-then rules are used, in our case, for revealing a set of If-and-only-if rules, as well as rules having several conditions that are unexpected relative to simpler rules.

The presented data-mining algorithm is proven to reveal –

1. All the IF-THEN rules that meet user-predefined thresholds.

2. All the IF-THEN-NOT rules (i.e., if the condition holds the result does not hold)

3. A set of optimal IF-AND-ONLY-IF rules (i.e., necessary and sufficient conditions)

Accordingly, it calculates the confidence level (i.e. error probability or significant level) for each rule. The algorithm analyses the UNEXPECTED RULES, presenting interesting and rare cases, which might be very important in some of the Biology applications.

The algorithm has been applied successfully, to different areas in life sciences. We shall present two highly diverse applications:

- The analysis of anthropometrical and blood screening data for ethnic group classification (joint work with Bat-Sheva Bonne-Tamir)
- The analysis of small-molecule binding-profiles for applications in nanobiotechnology and biochemistry (joint work with Ori Kalid and Amihay Freeman). The data-mining algorithm was used to define the necessary and sufficient conditions for a cavity on the molecular surface to bind a small molecule. This application was used as part of a screening system that selects protein building blocks for nanostructures with pre-designed geometry.

e-mail: miriam@post.tau.ac.il

Algorithm and Complexity for Designing Multi Route Synthesis

Akavia, A.[1], Aronowitz, H.[2], Lerner, A.[1], Senderowitz, H.[3] and Shamir, R.[1]
[1] School of Computer Science, Tel Aviv University
[2] Intel
[3] Peptor LTD.

It is widely accepted that success of lead identification and optimization is greatly enhanced by synthesizing and screening sets of compounds, which best represent the property space relevant to the biological activity of interest. The two most widely used synthesis schemes are parallel and combinatorial synthesis. Parallel synthesis enables cherry picked sets of compounds to be synthesized. Such sets, if properly selected, adequately represent the property space but are limited in size since each compound is synthesized individually. Combinatorial synthesis enables the production of large sets of molecules in a relatively short time but provides a poorer representation of the property space. The more complex multi route mix & split scheme, introduced by Cohen and Skiena, is a compromise between the parallel and the combinatorial procedures. It combines the advantages of both methods by enabling the rapid synthesis of fairly large sets of compounds while providing a better representation of the property space due to the relaxation of the combinatorial constraint. We propose an algorithm for the design of complex, multi route mix & split schemes that enable the synthesis of as many of the desired set of compounds as possible, within given resource constraints (*i.e.*, number of colons per step and total number of beads). We generate a graph representation of the synthesis procedure and perform a search for the optimal graph subject to the resource constraints. Each graph is assigned a score based on the number of the desired compounds it produces. We demonstrate the application of the algorithm on sets of compounds selected from a pre-defined property space. This scheme may be combined with diversity or similarity selection methods to provide adequate representation of the property space under given synthesis constraints.

We also explored the complexity of designing multi route mix & split schemes for a given set of desired compounds. Using computational complexity theory tools, we established NP-Completeness and hardness of approximation for several variants of the synthesis design problem.

A New Branch and Bound Feature Selection Algorithm

Frank, A.,[1] Geiger, D.[1] and Yakhini, Z.[1,2]
[1] Department of Computer Science, Technion
[2] Agilent Laboratories, Israel

Feature selection is an essential step to enhance correct classification in the presence of many irrelevant features and a small number of samples. For example gene expression data contains thousands of genes per sample, often with only a few dozens of samples. Most genes measured in a DNA microarray assay are irrelevant to the classification task. Identifying the few genes that affect classification is the task of feature selection.

The efficient application of expression profiling as a diagnostic tool [1] is highly dependent on making decisions based on reasonably small numbers of genes. Simple assays will be more cost-effective and much more robust. Enabling decisions based on a small number of parameters will also allow for building redundancy and control measurements into the process. Thus feature selection in the context of expression data may be key to the development of Pharmacogenomics.

We present herein an approach that drastically reduces the number of different feature subsets that need to be evaluated for realistic data including gene expression data. Using bounds on the Bayesian classification error, derived from monotonic additive distance measures such as the *Bhattacharyya distance* [2], our algorithm prunes subsets of features that are no longer candidates for having the lowest error, given the subsets examined so far. When the computational savings are not sufficient, we augment our approach with a preprocessing greedy identification of a sufficiently small subset of the most promising features and use this subset as the input to the main algorithm.

The algorithm we present is a *Branch and Bound* algorithm. This approach can be described as an exploration of a state space tree for the problem at hand. At each point of searching the tree, a bound is computed of the best solution possible in the current subtree. Promising nodes in the tree are expanded, whereas nodes for which the lower bound is larger than the best solution found so far, are pruned.

We tested our algorithm on three gene expression datasets: leukemia data [3], breast cancer [4] and prostate cancer [5]. Our algorithm selected subsets of up to 5 genes from each dataset, usually pruning more than 90% of the subsets in the process. The selection was done from the 100 genes having the highest Bhattacharyya distance. Using the gene subsets selected by our algorithm, a naive Bayesian classifier [2] has shown high classification success rates.

In the leukemia dataset, none of the 5 genes selected by our algorithm, are in the list of 50 discriminating genes used in [3], in which they had 5 misclassifications. We classified the entire test set correctly. In the breast cancer dataset, using a subset of 5 genes selected by our algorithm, we managed to classify correctly all the samples in the test set. Of these 5 genes, the first three appear in the list of 50 genes described in [4] (ranked 1, 12, and 27, respectively), while the last two do not. Perfect classification of the test set was also achieved by this work using all 3389 genes (which are reduced by PCA to 10 components). In the prostate cancer data, of the 5 most common genes from the subsets chosen in a LOOCV experiment for selecting subsets of 3 genes, the first two are amongst the top 100 genes mentioned in [5] (ranked 53 and 1, respectively). Using these 5 genes, our algorithm classified correctly 98% of the samples in a LOOCV test, higher than the any of the models using 4 to 256 genes mentioned in [5].

It is worthy to emphasize that our algorithm selects very small subsets, as small as 3-5 genes, compared to up to 50 genes or more needed in some of other approaches described in the literature. After selecting the gene subsets from microarray data consisting of thousands of genes, it might be possible to use these small subsets of genes for tissue classification employing small-scale low-cost gene expression measurement methods, such as RT-PCR. The development of assays based on profiling small sets of genes is crucial to cost-effectiveness of such methods in clinical practice.

References

[1] Bittner, M., *et al.* (2000), Nature, 406(13), p.536-540.

[2] Fukunaga, K., (1990), Introduction to Statistical Pattern Recognition, Academic Press.

[3] Golub T., *et al.* (1999), Science, 286, p. 531-537.

[4] Gruvberger, S., *et al.* (2001), Cancer Research, 16, p. 5979-5984.

[5] Singh, D., (2002), Cancer Cell, 1(2), p.203-209.

Finding Approximate Tandem Repeats in Genomic Sequences

Wexler, Y.[1], Kashi, Y.[2] and Geiger, D.[1]
[1] Department of Computer Science, Technion
[2] Department of Biotechnology, Technion

Genomic sequences tend to contain consecutive copies of patterns known as *tandem repeats* (TR). These occur due to DNA repair systems and other mechanisms, not all fully understood. Tandem repeats have proven useful as markers in genetic analysis due to the high degree of polymorphism observed in their number (e.g. for DNA fingerprinting applications [2]). Sometimes, such repeats are known to be the cause of a disease like in the case of Fragile-X syndrome and Huntington's disease.

A perfect tandem repeat is defined as a string of nucleotides, which is repeated consecutively at least twice. Many $O(n \cdot \log n)$ algorithms have been presented for finding perfect tandem repeats. However, in practice, mutations, translocations, and other biological events render the copies imperfect. This often results in *approximate tandem repeats* (ATR) defined as a string of nucleotides repeated consecutively at least twice such that, with sufficiently high probability, they originated from the same string. Finding ATRs in a genome is clearly a harder task than finding perfect repeats and has been addressed several times. The IBM bioinformatics group, have presented the best results so far [3] by altering their own TEIRESIAS pattern-finding algorithm.

The algorithm we present herein has *screening* and *verification* stages. In the screening stage, the possibility is evaluated of finding an ATR of certain length, at a certain position, with sufficiently high probability. In the *verification* stage, an alignment for each resulting candidate is generated and subsequently accepted or rejected as an ATR according to statistical criteria.

The *screening stage* uses a statistical model in which we consider matching two adjacent copies of a pattern of length t as a sequence of t independent Bernoulli trials. We replace the search for k contiguous successful nucleotide matches suggested by Benson [1] with a less stringent approach, as follows.

Consider a comparison to be a series of w independent Bernoulli trials, where w is a function of the pattern length t. Each trial compares two aligned nucleotides and the comparison passes if at least $l(w)$ such trials succeed. Three comparisons of length w are performed for each starting offset i, where $0 \le i \le t - w(t)$, to account for a single insertion, deletion or no-change at location i. We define the success list $S_t(i)$ of i with respect to the pattern length t to be the set of integers $j \in [i, i + t - w]$ for which the comparison starting at position j succeeded. The cardinality of $S_t(i)$ is called the score of i with respect to t.

In order for a position to pass the screening and become a candidate for verification, it has to satisfy several statistical criteria regarding the P-value of its score and a *valid* distribution of successful comparisons. We perform these comparisons on an entire genome of size n for all pattern lengths up to some $T_{\max}$ in time $O(T_{\max} \cdot n)$. This complexity is achieved because, instead of computing each comparison separately, they are computed incrementally while sliding over the genome.

The *verification stage*, when given a candidate ATR starting at position i with pattern length t, generates an alignment between the pattern starting at position i and the one starting at position $i+t$, both of length t, using a predetermined scoring system. If the alignment score is over a threshold, which depends on the scoring system, then the candidate is accepted as an ATR.

Results

The algorithm presented herein, aside of having a linear time complexity in genome size n when fixing $T_{\max}$, also performs well on real examples. When running over the yeast chromosome 1, which is the same genomic sequence used by IBM bioinformatics group to establish their results, using the same scoring system for verification as they did, and finding ATRs with length ranging from 10 to 300, the following facts emerge:

- Our algorithm found more than twice as many ATRs as was previously reported, finding also all the ones known before.
- A low rate of false positive candidates occurs during the screening stage (~3%). Consequently, the running time was less than 1% of the time reported by the IBM group, on a machine with no greater power.
- Our algorithm is parameter-free and does not require additional input from the user, aside of biological data, sequences, and a desired scoring system for alignment.

References

1. Benson, G. 1998. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acid Research*, vol. 27 pp. 573-580.

2. Inman, K. and Rudin, N. 1997. An introduction to forensic DNA analysis. {\em CRC press,} Boca Raton, Florida.

3. Stolovitzky, G., Gao, Y., Floratos, A. and Rigoutsos, I. 1999. Tandem repeat detection using pattern discovery, with applications to identification of yeast satellites. {\em IBM T.J.Watson Research Center.}

e-mail: ywex@cs.technion.ac.il
e-mail: kashi@tx.technion.ac.il
e-mail: dang@cs.technion.ac.il

Fast Implementation of e-PCR on Cray SV1 System

Tal, J.[1], Lapidot, M.[2] and Safran, M.[3]

[1] Cray LTD., Herzliya, Israel

[2] Department of Molecular Genetics, Weizmann Institute of Science

[3] Bioinformatics and Biological Computing Unit, Biological Services, Weizmann Institute of Science

The highly specific and sensitive PCR process provides the basis for sequence-tagged sites (STSs), unique landmarks that have been used widely in the construction of genetic and physical maps of the human genome. Electronic PCR (e-PCR) refers to the process of recovering these unique sites in DNA sequences by searching for subsequences that closely match the PCR primers and have the correct order.

The Weizmann Institute's GeneCards/UDB project uses the e-PCR program from NCBI to search for known STS primer sequences in contigs, in order to produce a more precise positioning of markers, genes, and EST clusters in sequenced regions.

The standard implementation of e-PCR on a Sun/Solaris platform takes a few days to complete the search for primers sequences within the whole human genome.

A fast approach to e-PCR was implemented on a Cray-SV1 Supercomputer. The Cray SV1 Parallel Vector architecture has special hardware features that are used to boost the performance of the STS search procedure by an order of magnitude.

Performance of sample runs shows a boost of ~8X in cpu performance between our fast approach on Cray SV1@300MHz and SGI origin2000@400MHz, when no mismatches are allowed. This ratio grows when mismatches are allowed. When running with a threshold of two mismatches a ratio of ~28X was measured. The performance ratio of the new Cray code relatively to the standard code running on the Cray is ~7X with no mismatches and ~13X with two mismatches.

Many Needles in a Haystack: Finding the Global Optimum and Best Populations in Biomolecular Systems

Rayan, A., Glick, M., Gorelik, B., Noy, E., Brinker, G., and Goldblum, A.
Department of Medicinal Chemistry and Natural Products, School of Pharmacy,
Hebrew University of Jerusalem

A new search algorithm that finds global minima and best populations in complex combinatorial problems has been successfully applied to a few problems in biomolecular structure, such as large loop predictions and homology modeling, side chain positioning, predicting proton positions from crystallographic results and cyclic peptide conformations. It is currently applied to flexible protein-ligand docking, flexible protein-protein interactions, molecular conformations, and to various aspects of structure based drug design.

The above have been achieved by applying a general approach to searching complex surfaces of cost functions. An initial stochastic construction of a system's configuration is followed by eliminating values that consistently contribute to maximal cost function values. Further iterations reduce the overall number of values to a point from which a full exhaustive search is performed and "best populations" may be constructed. Such populations are relevant to structure and function analysis. Comparison between stochastic and exhaustive searches prove that this new search reproduces the real populations in the problems studied. The efficiency of the algorithm encourages parallel processing for substantial increase in problem size.

We will use the "traveling salesman problem" to demonstrate how this algorithm may be applied to a well known difficult issue which has an exact cost function. Energy functions in structural biology are huge approximations and still lack the ability to "lead" to the best answers. This algorithm may however be employed also for finding improved cost functions to problems of structural biology.

References:

1. M. Glick and A. Goldblum, Proteins 38, 273-287 (2000)
2. M. Glick, A. Rayan and A. Goldblum, PNAS 99, 703-8 (2002)
3. M. Glick and A. Goldblum, Patent request PCT WO0139098 (2001), http://ep.espacenet.com/

Automating the Superparamagnetic Clustering Method

Agrawal, H. and Domany, E.
Department of Physics of Complex Systems, Weizmann Institute of Science

Superparamagnetic clustering is a method for clustering data that exploits the phase transitions in grannular ferromagnets for solving the clustering problem. Each data point is associated with a spin; the mapping from the clustering problem to a ferromagnet is dependent on the value of a parameter ‘K’, which controls the number of neighbors with which each spin interacts. The value of K determines the kind of highly inhomogeneous lattice to which the data are transformed, and the solution of the clustering process exhibits non-trivial dependence on this parameter. Untill recently K was determined by exploring a wide interval of possible values, which is a computationally expensive procedure. We present a method for determining the range of this parameter for which best clustering solutions are obtained. The method is fully automated and gives the optimal range of ‘K’ almost instantaneously.

GeneCards 3.0: An Object-Oriented Approach

Safran, M.[2], Shen-Orr, S.[3], Solomon, I.[2], Lapidot, M.[1], Shmueli, O.[1], Rosen, N.[1], Adato, A.[1], Ben-Dor, U., Esterman, N., Chalifa-Caspi, V.[2], and Lancet, D.[1]

[1] Dept. of Molecular Genetics, [2] Bioinformatics and Biological Computing Unit, Biological Services, [3] Dept. of Molecular Cell Biology, Weizmann Institute of Science

GeneCards is a database of human genes, maps, proteins, and diseases, with associated software that retrieves, integrates, and displays gene centered human genome information [1,2]. Versions 2.*xx* have stressed features and usability, including query reformulation and grappling with comprehensiveness versus compactness, in order to present *just the right mix* of detail and hyperlinks. GeneCards has gained widespread popularity, as evidenced by over two million hits at the home site, mirroring by 25 academic institutions around the world and ever-growing commercial interest. Version 3.0 strives to maintain its successful look and feel, data-mining heuristics, feature enhancements and data upgrades, while strengthening the infrastructure, and standardizing data formats using object-oriented and XML (**E**xtensible **M**arkup **L**anguage[3] technologies.

We present the pros and cons of using object-oriented Perl[4] and our hybrid approach of implementing an object-oriented skeleton with some non-object-oriented internals to enhance the system's efficiency.

XML[3] is a meta-language that supports customized tags for describing and providing semantic meaning to structured data. This open and self-describing format can be easily parsed by other applications and its typed elements can be arranged within others to form a nested hierarchy. We present two XML schemas for representing the GeneCards data, *GeneCardByResource* and *GeneCardByFunction* and their impact on the GeneCards display software. Moving to XML will also facilitate the implementation of an expanded search engine beyond the current text-based capability to include context-specific searches.

References:

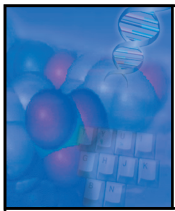
1. Rebhan M, Chalifa-Caspi V, Prilusky J and Lancet D, *"GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support"* Bioinformatics, 14(8):656-664 (1998).

2. Safran M *et al, "GeneCards 2002: towards a complete, object-oriented human gene compendium"* , Bioinformatics, **in press.**

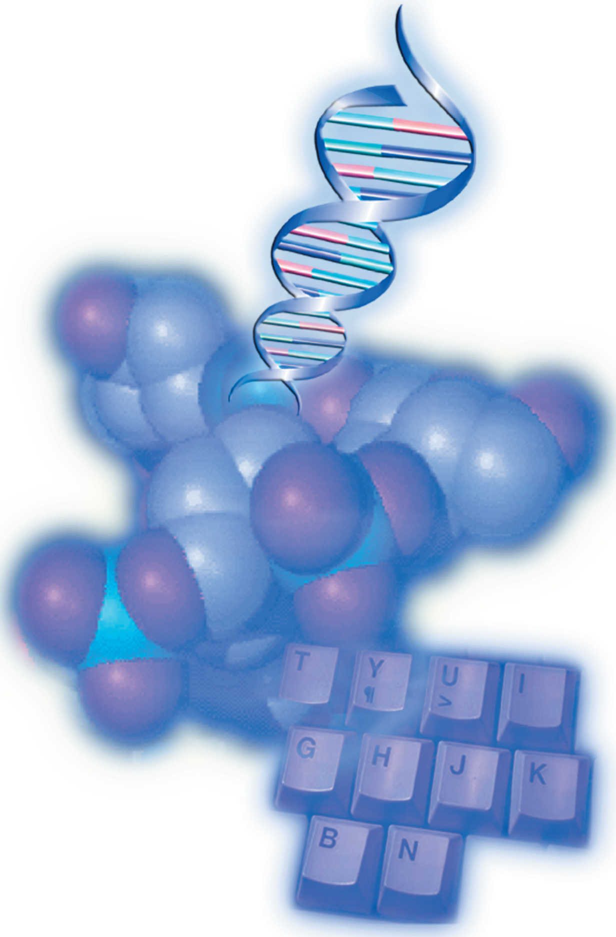
3. http://www.w3.org/XML

4. Conway Damian - *Object Oriented Perl*, Manning Publications Co. 1st edition, 2000

e-mail: cards@bioinformatics.weizmann.ac.il
website: http://bioinformatics.weizmann.ac.il/cards/



Workshop Abstracts



Genomics & Medical Bioinformatics

Jacques S. Beckmann, Chair

A New Molecular Approach for Haplotyping in Large Population-Based Association Studies

Benjamin Yakir, Dept. of Statistics, Hebrew University of Jerusalem

Authors: Ester Inbar, Benjamin Yakir and Ariel Darvasi, Hebrew University of Jerusalem

Determination of haplotype frequencies (the joint distribution of genetic markers) in large population samples is a powerful tool for association studies.

Population haplotype frequencies evaluate linkage disequilibrium between markers. Haplotypes are of great value for association studies due to their greater extent of variability. Therefore, a single haplotype may capture any given functional polymorphism with higher statistical power than its SNP components. The statistical estimation of haplotype frequencies, usually employed in LD studies, requires the individual genotyping for each SNP in the haplotype, thus making it an expensive process.

In this talk, we describe a new method for direct measurement of haplotype frequencies in DNA pools, by allele-specific, long-range, amplification of the pool.

The proposed method allows high throughput genotyping of haplotypes composed of two SNPs in close vicinity (up to 20Kb).

We will discuss some of the statistical implications of applying this approach in large population based association studies.

Expression Profiling and the Quest for BRCAx

Zohar Yakhini, Agilent Technologies and Technion

Recent studies on molecular level classification of cancer cells produced results that strongly indicate the potential of gene expression profiling assays as diagnostic and segmentation tools and as a basis to the discovery of putative disease subtypes. At Agilent labs we are developing measurement and data analysis techniques that enable the acceleration of accurate and statistically sound expression based studies.

I will discuss several aspects of designing expression profiling assays and of analyzing resulting data. In particular I shall describe methods for un-supervised class discovery. In classified gene expression data an overabundance of genes that sharply separate the classes is typically observed. Reversing the relationship between overabundance and biological meaning we developed an approach to unsupervised class discovery. The output of the latter computational task is a partition of the set of samples which is supported by a statistically meaningful set of genes. Overabundance analysis is used as a figure of merit in our class discovery approach.

The genetic determinants of familial breast cancer that is not attributable to BRCA1/2 mutations are generally designated as BRCAx. I will discuss the application of class discovery and other data analysis tools in a BRCAx study recently completed with NHGRI collaborators.

Understanding the Emergence of Specific Information in Living Systems

Irun Cohen, Dept. of Immunology, Weizmann Institute of Science

Living systems *mine* information from their environments and *create* information within themselves; that's how they manage to live. Our task is to understand how *they* and *we* do it.

Proteomics & Bioinformatics

Meir Edelman, Chair, Vladimir Sobolev, Co-chair

The workshop will consider the issue of “Accuracy of protein structure determination by X-ray crystallography and molecular modeling”. Are the existing experimental and/or modeling tools sufficient to pinpoint the interactions that stabilize a protein or its complexes? What portion of the PDB structures is independently resolved? How accurate are current structure predictions? The workshop will include panel speakers as well as audience participation.

Panel:

Daniel Fischer: Dept. of Computer Science, Ben Gurion University
Felix Frolow: Dept. of Molecular Microbiology and Biotechnology, Tel Aviv University
Vladimir Sobolev: Dept. of Plant Sciences, Weizmann Institute of Science

Intellectual Property & Bioinformatics

Jonathan Gressel, Chair

Protecting your Crystal, Crystallography Data and their Uses

Gal Ehrlich, Patent Attorney, e-mail: gal@ipatent.co.il

The structural data of proteins and their association with ligands obtained via X ray crystallography is highly valuable and useful for drug development. The talk will focus on patenting strategies directed at protecting the structural data and information emerging from X ray crystallography studies. Certain subject matter aspects pertaining to X ray crystallography patents, for which there is an open debate with respect to patentability will also be discussed.

A Proposed System for ‘Biobarcoding’^{TM} Organisms

Jonathan Gressel, Department of Plant Sciences, Weizmann Institute of Science

These are a variety of needs for devising simpler recognition methods for organisms marketed in commerce or released in the environment; whether they are conventionally selected, mutant, or transgenic bacteria, fungi, plants or animals. The needs include:

The need for protection for patented or other IP lines, where IP takes on either designation: “Intellectual Property” or “Identity Preserved”. It is often hard to prove that a line has been ‘miss-appropriated’ by a competitor.

Labeling Regulatory authorities and various consumer groups are demanding labeling of transgenic commodities. They spend vast sums typically probing for commonly used promoters or selectable marker genes and not for the trait genes, in an effort to save. Even when transgenics are discovered by such ‘kits’, there is no information as to source. Thus, regulatory authorities may wish to consider simple, common recognition sequences for detecting transgenic or other organisms in the market place.

The need to trace organisms in the environment. The use of mycoherbicides and other live organisms as inoculants and or as biocontrol agents to control weed, bacterial, fungal, or insect pests is increasing. This need is irrespective of whether indigenous or transgenic. Many of the agents are closely related to known pathogens or pests and there are claims that an organism may change its host range and attack valuable species. There are complicated DNA fingerprinting techniques to accurately ascertain causality, but they cannot be used to probe what released organism might be present. There are also fears that organisms will mutate or introgress with other organisms, and there are needs to know whether the organism changed host range (with consequences of liability) or whether an epidemic was due to wild strains. These issues will become more acute with transgenically-enhanced organisms.

The simplest detection system for differentiating a large number of products is the “bar code” system. A simple genetic analogy encoded in DNA sequences – “biobarcodes” is proposed. A set of two universal 'nonsense' (non-coding) nucleotide sequences is designed. These can be detected by a set of universal PCR primers that can be used to recognize all biobarcodes. The universal primers are long enough that a few mutational changes in the initial universal sequence will still allow it to be recognized by a PCR primer. The universal recognition codes are followed by a designed and assigned nonsense sequence that is long enough to allow tens of millions of different such sequences to be generated, and again allow for some mutational changes. Neither the initial universal recognition sequence nor the particular individual strain sequence should even vaguely resemble nonsense sequences reported in any gene data base. The algorithms used to generate the sequence are designed to exclude sequences that could self anneal, preventing the taq polymerase from amplifying the DNA. Frame shift mutations should not render any part of the biobarcode sequence as an open reading frame coding for a peptide – stop codons are interspersed so as to prevent frameshift mutations to form long open reading frames. The biobarcodes should be assigned by a single assigner, and the assigned codes are to be publicly available. The biobarcode DNA can be co-transformed with the gene of choice. In other cases, an excisable selectable marker will be needed, so that just the bar code remains after transformation.

The PCR amplified barcodes can be automatically sequenced and compared to the barcode database to ascertain the source of the organism. Should there be a possibility of introgression of the barcode from the initial organism into another strain or species, R or AFLP can be used to further verify the source

Biology Applications for Computational & Mathematical Methods

Eytan Domany, Chair

Controlling the False Discovery Rate in Behavioral Genetics and Microarrays Analysis

Yoav Benjamini, Dept. of Statistics and Operations Research, Tel Aviv University

Any study based on statistical evidence is prone to produce false discoveries. Testing statistical hypotheses at the traditional 0.05 level is a way to limit the probability of producing such a false discovery when it is not real. Alas when many such tests are conducted in a study, the probability of producing a false discovery increases dramatically. On the other hand, limiting this probability in the traditional way incurs deterioration in the probability of detecting true discoveries.

The control of the false discovery rate (FDR) has been suggested as an intermediate approach to the problem. Procedures that control the FDR at a desired level are gaining popularity in areas of science where the problems encountered are large. Two such areas will be discussed in this talk – behavioral genetics and gene expression microarray data. The two will serve as a vehicle to discuss the FDR approach, the simple procedures that may be used to control it, and their properties.

Large-Scale Clustering of Protein Sequences: Some Theory and Some Practice

Nati Linial, Dept. of Computer Science, Hebrew University of Jerusalem

In this presentation, I will review some of our recent work on the large-scale classification and analysis of proteins. I will explain some of the algorithmic concepts that underlie our work. Among the algorithmic tools that we employ are: Spectral data analysis, combinatorial algorithms to detect significant domains and small distortion metric embeddings.

Cluster Analysis of Gene Expression Data

Gad Getz, Dept. of Physics and Complex Systems, Weizmann Institute of Science

A single microarray experiment allows simultaneous measurement of the expression level of thousands of genes. A typical experiment uses a few tens of such microarrays, each focusing on one sample - such as material extracted from a particular tumor. Hence the results of such an experiment contain several hundred thousand numbers, that come in the form of a table, of several thousand rows (one for each gene) and 50 - 100 columns (one for each sample).

We developed and applied a “data mining” method, called Coupled Two-Way Clustering} (CTWC), to extract biologically relevant data from such matrices. By an iterative clustering procedure the method reveals correlations that involve small subsets of genes and samples. The method can identify sets of correlated genes which usually belong to the same biological process or pathway, and uncover types and sub-types of diseases. The algorithm is designed to link cellular conditions to their relevant genes by finding conditional correlation among them. I will present results obtained from analyses of several types of cancer.

The CTWC method was applied by others and by us to numerous data types including gene expression data, antigen reactivity data, analysis of sugar compounds, document categorization and low-temperature phases of short range spin glasses.

Sponsors

Academic Sponsors at the Weizmann Institute of Science

The Maurice and Gabriela Goldschleger Conference Foundation at the Weizmann Institute of Science

The Kimmelman Center of Biomolecular Structure and Assembly

The Crown Human Genome Center

The Arthur and Rochelle Belfer Institute of Mathematics and Computer Science

Faculties of the Life Sciences

INN - The Israeli National Node

Commercial Sponsors

ATLAS WORLD LTD.
ATLAS world

CRAY

IBM יבמ

sgi
One step ahead

Spotfire®

פארמה ג'וב
שרותי השמה ואיבחון פארמצבטיים

Author Index

Scientific Committee:

Shmuel Pietrokovski, Hanah Margalit,
Daniel Fischer, Haim Wolfson

Local Committee:

Shmuel Pietrokovski, Chair
Ron Pinter, Co-chair
Joel Sussman, Marvin Edelman,
Eitan Rubin
e-mail: bioinfo@weizmann.ac.il
http://bioinfo.weizmann.ac.il/aibs02/

Graphics Dept. ▶ Weizmann Institute of Science ◆ Graphics Design: Genia Brodsky