(Sackler NAS Colloquium) Frontiers of Bioinformatics: Unsolved Problems and Challenges

Organized by Samuel Karlin, David Eisenberg and Russ Altman, National Academy of Sciences

ISBN: 0-309-65400-9, 58 pages, 8 1/2 x 11, (2005)

Frontiers of Bioinformatics: Unsolved Problems and Challenges

National Academy of Sciences
Washington, DC

Arthur M. Sackler, M.D.
1913–1987

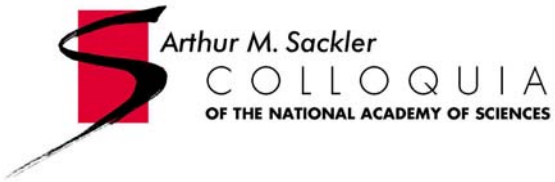
Born in Brooklyn, New York, Arthur M. Sackler was educated in the arts, sciences, and humanities at New York University. These interests remained the focus of his life, as he became widely known as a scientist, art collector, and philanthropist, endowing institutions of learning and culture throughout the world.

He felt that his fundamental role was as a doctor, a vocation he decided upon at the age of four. After completing his internship and service as house physician at Lincoln Hospital in New York City, he became a resident in psychiatry at Creedmoor State Hospital. There, in the 1940s, he started research that resulted in more than 150 papers in neuroendocrinology, psychiatry, and experimental medicine. He considered his scientific research in the metabolic basis of schizophrenia his most significant contribution to science and served as editor of the *Journal of Clinical and Experimental Psychobiology* from 1950 to 1962. In 1960 he started publication of *Medical Tribune*, a weekly medical newspaper that reached over one million readers in 20 countries. He established the Laboratories for Therapeutic Research in 1938, a facility in New York for basic research that he directed until 1983.

As a generous benefactor to the causes of medicine and basic science, Arthur Sackler built and contributed to a wide range of scientific institutions: the Sackler School of Medicine established in 1972 at Tel Aviv University, Tel Aviv, Israel; the Sackler Institute of Graduate Biomedical Science at New York University, founded in 1980; the Arthur M. Sackler Science Center dedicated in 1985 at Clark University, Worcester, Massachusetts; and the Sackler School of Graduate Biomedical Sciences, established in 1980, and the Arthur M. Sackler Center for Health Communications, established in 1986, both at Tufts University, Boston, Massachusetts.

His pre-eminence in the art world is already legendary. According to his wife Jillian, one of his favorite relaxations was to visit museums and art galleries and pick out great pieces others had overlooked. His interest in art is reflected in his philanthropy; he endowed galleries at the Metropolitan Museum of Art and Princeton University, a museum at Harvard University, and the Arthur M. Sackler Gallery of Asian Art in Washington, DC. True to his oft-stated determination to create bridges between peoples, he offered to build a teaching museum in China, which Jillian made possible after his death, and in 1993 opened the Arthur M. Sackler Museum of Art and Archaeology at Peking University in Beijing.

In a world that often sees science and art as two separate cultures, Arthur Sackler saw them as inextricably related. In a speech given at the State University of New York at Stony Brook, *Some reflections on the arts, sciences and humanities,* a year before his death, he observed: ''Communication is, for me, the *primum movens* of all culture. In the arts. . . I find the emotional component most moving. In science, it is the intellectual content. Both are deeply interlinked in the humanities.'' The Arthur M. Sackler Colloquia at the National Academy of Sciences pay tribute to this faith in communication as the prime mover of knowledge and culture.



FRONTIERS OF BIOINFORMATICS: UNSOLVED PROBLEMS AND CHALLENGES

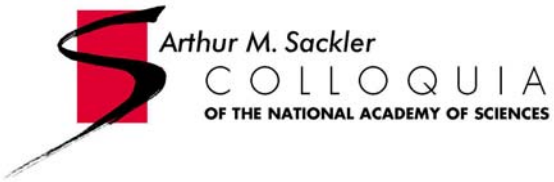
OCTOBER 15-17, 2004

The Beckman Center, Irvine, CA

Organized by Samuel Karlin, David Eisenberg and Russ Altman

Table of Contents

Cover Art courtesy of Samuel Karlin, Stanford University



FRONTIERS OF BIOINFORMATICS: UNSOLVED PROBLEMS AND CHALLENGES

October 15-17, 2004
Beckman Center of the National Academies
100 Academy Drive, Auditorium
Irvine, California

Organized by Samuel Karlin, David Eisenberg and Russ Altman

PROGRAM

Friday, October 15

7:45 pm	Buses Depart Hyatt Newporter for Beckman Center
8:00-10:00 pm	Registration, Welcome Reception, and Poster Session
10:00 pm	Buses Depart Beckman Center for Hyatt Newporter

Saturday, October 16

7:15 am and	
7:45 am	Buses Depart Hyatt Newporter for Beckman Center
7:30 am	Breakfast

Opening Comments

8:30 am	Samuel Karlin (Stanford University)

Session I: Informatics of the Human Genome (8:35 am – 12:10 pm)

Chair	Samuel Karlin (Stanford University)
8:35 am	George Miklos (Secure Genetics Pty Limited and Human Genetic Signatures Pty Limited), *Clinical Challenges for Bioinformatics*
9:20 am	Mark Gerstein (Yale University), *Human Genome Annotation*
10:05 am	Break
10:35 am	David Haussler (University of California, Santa Cruz), *Using Evolution to Explore the Human Genome*
11:20 am	Pavel Pevzner (University of California, San Diego), *Transforming Men into Mice (and into Cats, Dogs, Cows, Rats, Chimpanzees, etc.): Evolutionary Lessons from Mammalian Sequencing and Comparative Mapping Projects*
12:10 pm	Lunch

Session II: Motifs and Genomics (1:30 – 3:00 pm)

Chair	Russ Altman (Stanford University)
1:30 pm	Peer Bork (European Molecular Biology Laboratory), *Genome Evolution and Protein Networks*
2:15 pm	Phil Green (Howard Hughes Medical Institute and University of Washington), *Signal and Noise in Genomic Sequences*
3:00 pm	Break

Session III: Protein-Protein Interactions (3:30 – 6:00 pm)

Chair	Valerie Daggett (University of Washington)
3:30 pm	David Eisenberg (University of California, Los Angeles), *Protein Interactions*
4:15 pm	Hanah Margalit (The Hebrew University of Jerusalem)*, From Cellular Networks to Molecular Interactions and Back*
5:00 pm	Shoshana Wodak (Hospital for Sick Children), *Protein-Protein Interactions: The Challenge of Predicting Specificity*
6:15	Reception and Poster Session
8:00	Dinner Russell F. Doolittle (University of California, San Diego), *Regarding Irreducible Complexities*, Introduced by David Eisenberg (University of California, Los Angeles)
10:30	Buses Depart Beckman Center for Hyatt Newporter

Sunday, October 17

7:15 am and 7:45 am	Buses Depart Hyatt Newporter for Beckman Center
7:30 am	Breakfast

Session IV: Regulation with RNA and Aspects of Splicing (8:30 – 10:45 am)

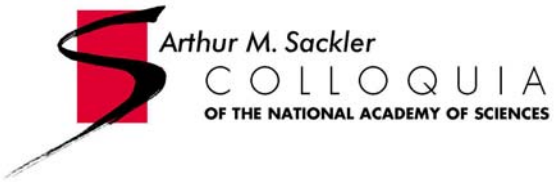
Chair	David Eisenberg (University of California, Los Angeles)
8:30 am	Sean Eddy (Washington University, St. Louis), *The Modern RNA World: Computational Screens for Noncoding RNA Genes*
9:15 am	Christopher Burge (Massachusetts Insttitue of Technology), *Toward an RNA Splicing Code*
10:00 am	Christopher Lee (University of California, Los Angeles), *Discovering Evolutionary Mechanisms from Multiple Metrics of Molecular Evolution*
10:45 am	Break

Session V: Protein Structure (11:00 am – 12:30 pm)

Chair	George Miklos (Secure Genetics Pty Limited and Human Genetic Signatures Pty Limited)
11:00 am	Helen Berman (Rutgers University), *Probing the PDB*
11:45 am	Michael Levitt (Stanford University), *Structural Alignment and Classification of all Known Protein Structure*
12:30 pm	Lunch

Session VI: Transcription and Translation in Eukaryotic Genomes (1:30 – 4:45 pm)

Chair	George Miklos (Secure Genetics Pty Limited and Human Genetic Signatures Pty Limited)
1:30 pm	Volker Brendel (Iowa State University), *Comparative Plant Genomics: Evaluation of the Model Genome Concept*
2:15 pm	Terry Gaasterland (Rockefeller University and University of California, San Diego/SIO), *Lessons from the Arabadopsis Genome: Decoding Evidence for Novel Transcription*
3:00 pm	Break
3:15 pm	Russ Altman (Stanford University), *Building Genotype Phenotype Data Resources*
4:00 pm	Samuel Karlin (Stanford University), *Highly Expressed Genes Based on Codon Usage Biases in Archaeal and Eukaryotic Genomes*
5:00 pm	Buses Depart Beckman Center for Hyatt Newporter and Orange County Airport



FRONTIERS OF BIOINFORMATICS: UNSOLVED PROBLEMS AND CHALLENGES

October 15-17, 2004

PRESENTATION

ABSTRACTS

This page is intentionally left blank.

Clinical Challenges for Bioinformatics

George L Gabor Miklos
Secure Genetics Pty Limited and
Human Genetic Signatures Pty Limited, Sydney, Australia

The clinical validation of mathematically and statistically rigorous bioinformatic models in the prognostic contexts of human diseases, and in the evaluation of methylation signatures, is a difficult endeavor. However, if transcriptomic and proteomic data are to significantly enhance therapeutic protocols, they must provide an improvement on current treatments, where for example, adjuvant systemic administration of anti-cancer drugs generally yields net gains in survival of only a few months, and where in the case of breast cancers, only 10% or so of patients benefit from such treatments.

Genome wide analyses, exemplified by microarrays, have become a backbone of molecular research, but difficulties are emerging in their applications to cancers and complex diseases (Ein-dor et al., 2004, *Bioinformatics*, *in press*; Miklos and Maleszka, 2004, *Nature Biotechnology*, 22, 615; Yeung et al., 2004, *Genome Biology*, 5, R48). For example, analyses of the *same* lung cancer data by *different* bioinformatic pipelines, implemented by computer experts, statisticians and bioinformaticians from academia and the pharmaceutical sector, have found almost no commonalities between the gene sets that are claimed to be of prognostic significance to patient survival (*Critical Assessment of Microarray Data Analysis* meeting 2003). In addition, these new analyses revealed little overlap with the genes that were considered to be of most importance in the original studies. A similar situation holds from microarray data on leukemias. The use of different commonly used preprocessing packages, such as MAS5.0, RMA and GCRMA, (Bumgarner, 2004), on data from smokers versus non-smokers, also yields largely non-overlapping gene sets. Furthermore, analysis of breast cancer survival datasets demonstrates that prognostic gene cohorts are not unique and that equally predictive lists can be produced from the same data. Thus, the “top” genes cannot be considered as the main candidates for anti-cancer treatment, since there are many different groups of “top” genes. Similarly, the genes that have been prioritized in neuropsychiatric disorders such as the schizophrenias, barely overlap with those that have emerged from clinical, *in situ*, SNP, drug perturbation, knockout and association studies. Finally, data from *Saccharomyces*, where genome-wide knockouts and phenotypic data have been compared to expression data, have shown that no simple relationship exists between genes selected on the basis of their expression level changes and the biology of the perturbed system (Birrell et al., 2002, *PNAS*, 99, 8778).

At the clinical level, continuing impediments to therapeutic progress are the ill-defined boundaries of most diseases at the level of the individual and the extensive phenotypic variation of human diseases; for example, different samples from the same tumor are molecularly highly heterogeneous. We are faced with broad categories such as the dementias, the cancers and AIDS, all of which are heterogeneous collections of perturbed biological systems that have undoubtedly reached their phenotypic endpoints by different trajectories in different individuals. Despite the implications of this heterogeneity, ultra-sensitive transcriptomic and proteomic technologies are nevertheless enthusiastically applied to human tissue samples, in many cases, inappropriately chosen ones. This is particularly dangerous when the etiology and the clinical symptoms are separated by decades, or when genomes, and hence cellular networks and attractor basins are massively imbalanced or rerouted owing to aneuploidogenic processes. A major challenge therefore, is to derive a robust mapping between the phenotypic space defined by physicians and the perturbed networks which occur at different levels from the molecular, developmental and neuroanatomical through to the cognitive. A second challenge involves deconvoluting the dynamics of change; how does the initial perturbation set in motion the

cascade of events which percolate and modify the various levels until some form of altered trajectory becomes clinically recognizable?

Finally, at the epigenetic level, each cell type has a different methylation signature that characterizes that cell type. However, there is a semi-fluid modulation of methylation signatures that characterizes each cell which is a result of all the epigenetic changes that have occurred since fertilization and the current tissue niche in which that cell resides. The unique cellular signature of an individual can alter owing to diet, age, stress, drugs and so forth. Hence, at the methylation coalface, we face a far more interesting and complex clinical situation than the current emphasis on hardware changes such as mutations, SNPs, and gross genomic imbalances, since methylation signatures are dynamic and context dependent. They provide snapshots of the *current network status* and hence of our current cellular operating systems. The rewards of rigorous bioinformatic analyses in this sphere are likely to be profound.

Human Genome Annotation

Z Zhang, P Harrison, Y Liu, N Carriero, D Zhang, P Bertone, J Karro, D Milburn, N Echols, J Rinn, M Snyder, M Gerstein
Yale University

A central problem for 21st century science will be the analysis and understanding of the human genome. My talk will be concerned with topics within this area, in particular annotating pseudogenes (protein fossils) in the genome. I will discuss a comprehensive pseudogene identification pipeline and storage database we have built. This has enabled use to identify >10K pseudogenes in the human and mouse genomes and analyze their distribution with respect to age, protein family, chromosomal location. One interesting finding is the large number of ribosomal pseudogenes in the human genome, with 80 functional ribosomal proteins giving rise to ~2,000 ribosomal protein pseudogenes. At end I will talk broadly about pseudogenes, in terms of their composition and mutation rates and I will compare pseudogenes in the human with those in a number of other model organisms, including worm, fly, yeast, and various prokaryotes. I will also talk about the problem of identifying pseudogenes in relation to the overall problem of finding genes in genome.

http://bioinfo.mbb.yale.edu
http://pseudogene.org

Comparative analysis of processed pseudogenes in the mouse and human genomes.
Z Zhang, N Carriero, M Gerstein (2004) Trends Genet 20: 62-7.

Identification of pseudogenes in the Drosophila melanogaster genome.
PM Harrison, D Milburn, Z Zhang, P Bertone, M Gerstein (2003) Nucleic Acids Res 31: 1033-7.

Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome.
Z Zhang, PM Harrison, Y Liu, M Gerstein (2003) Genome Res 13: 2541-58.

Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes.
Z Zhang, M Gerstein (2003) Nucleic Acids Res 31: 5338-48.

Studying genomes through the aeons: protein families, pseudogenes and proteome evolution.
PM Harrison, M Gerstein (2002) J Mol Biol 318: 1155-74.

Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome.
Z Zhang, P Harrison, M Gerstein (2002) Genome Res 12: 1466-82.

Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome.
PM Harrison, N Echols, MB Gerstein (2001) Nucleic Acids Res 29: 818-30.

Using Evolution to Explore the Human Genome

David Haussler
University of California, Santa Cruz

The reference sequence of the human genome was recently produced, along with drafts of the chimp, mouse, rat, dog, chicken and other genomes. Data and analysis of these is available on the genome browser at http://genome.ucsc.edu, a site that now averages more than 5,000 distinct users per day. This is the site where the first publicly accessible working draft of the human genome was posted. The site currently features an interactive “microscope” on the human genome and its evolution, via cross-species comparative genomics.

In 2002, a statistical estimate based on a simple measure of similarity between short orthologous segments in the human and mouse genomes suggested (very roughly) that about 5% of the human genome shows signs of being under purifying selection. Purifying selection occurs in the most important functional segments of the genome, where random mutations are mostly deleterious and hence are rejected by natural selection, leaving the orthologous segments in different species more similar than would be expected under a “neutral” mutation model. With more genomes now available, we find that this rough estimate of the fraction of the genome under purifying selection is holding up, and we are better able to pinpoint specific segments in the human genome that are evolving under this type of selection.

We are using new “context dependent” models of molecular evolution to find regions of the human genome that are not only under purifying selection, but are specifically evolving like protein-coding regions in genes. Here evolutionary analysis leads to a functional prediction. These methods have led to the prediction of many new human genes. Related analysis led to the discovery of a host of previously unexplored non-coding elements in the human genome that are under extremely strong purifying selection as well. We call these “ultraconserved” elements. Their function is currently unknown.

This work brings up an interesting information theoretic question: how well can we use the genomes of our present day animal relatives to reconstruct the evolutionary history of the individual bases of our genome, or, in other words, how much information about the ancestral state of our DNA bases was irrevocably lost? This depends a lot, but not exclusively, on how far back in time you want to go. Via simulation, we estimate that most of the DNA sequence of the common ancestor of all placental mammals can be predicted with 98% accuracy. This placental ancestral genome can in fact be reconstructed better than that of some more recent human ancestors because of the favorable phylogenetic tree topology present in the rapid radiation of species of placental mammals in the last part of the Cretaceous period. The full theory of the reconstructability of ancestral DNA bases, via a mutual information analysis, appears to be non-trivial.

References: see http://genome.ucsc.edu/goldenPath/pubs.html

Transforming Men into Mice (and into Cats, Dogs, Cows, Rats, Chimpanzees, etc): Evolutionary Lessons from Mammalian Sequencing and Comparative Mapping Projects

Pavel Pevzner
University of California at San Diego

In a pioneering paper, Nadeau and Taylor, 1984 estimated that surprisingly few genomic rearrangements have happened since the divergence of human and mouse 80 million years ago. Every genome rearrangement study involves solving a combinatorial puzzle to find a series of genome rearrangements to transform one genome into another. I will briefly describe some genome rearrangements algorithms and show how these algorithms shed light on previously unknown features of mammalian evolution. In particular, they provide evidence for extensive re-use of breakpoints from the same relatively short regions and reveal a great variability in the rate of micro-rearrangements along the genome. Our analysis also implies the existence of a large number of very short ``hidden'' synteny blocks that were invisible in comparative mapping data and were ignored in previous studies of chromosome evolution. These results suggest a new model of chromosome evolution that postulates that breakpoints are chosen from relatively short fragile regions that have much higher propensity for rearrangements than the rest of the genome.

This is a joint work with Glenn Tesler.

Genome Evolution and Protein Networks

Peer Bork
European Molecular Biology Laboratory, Heidelberg

Although with the availability of many completely sequenced metazoan genomes our understanding of functionality encoded therein increases, there are still numerous features in the genomes that need to be exploited for functional and evolutionary purposes. Here, I start off by describing an emerging, so far undescribed, new gene family in human that appears to drive the shaping of up to 10% of human chromosome II. Then I illustrate more generally the dynamics of gene content in metazoan genomes and how it correlates with various other measurements of genome evolution such as intron content, protein architecture or synteny. All of these measures indicate that the speed of evolution differs in some lineages. Over larger time scales some of the genomic features such as gene neighborhood indicate functional constraints that can be used for function prediction and for the construction of protein interaction networks with remarkable accuracy. Analysis of such networks reveals the functional modularity therein and how it changes in time.

Signal and Noise in Genome Sequences

Phil Green
Howard Hughes Medical Institute and University of Washington

Interpreting genome sequences requires distinguishing `signal', i.e. encoded functional elements, from `noise', i.e. non-functional, neutrally evolving sequence. We are working towards an improved understanding of both sides of this dichotomy.

The characteristics of non-functional sequence reflect underlying mutational processes, which remain poorly understood. Previous studies of mammalian pseudogene data by several investigators have revealed that transitions occur more frequently than transversions, G:C mutates to A:T at a higher rate than the reverse, and rates depend significantly on the flanking nucleotide context, with methylated C's in CpG dinucleotides being notable hotspots. Studies of synonymous coding substitutions have suggested a 'generation time effect' consistent with the idea that most mutations occur in conjunction with DNA replication. Recent work of Duret and others points to an important role for recombination in the substitution process, likely reflecting the effects of biased gene conversion.

Using data from the NISC project (www.nisc.nih.gov/), we have discovered (Nat Genet 33, 514-517 (2003)) a mutational asymmetry associated with transcribed regions that we believe reflects an asymmetry in DNA polymerase errors that is unmasked by transcription-coupled repair. This mutational asymmetry has acted over long evolutionary periods to produce a compositional asymmetry within transcribed regions of mammalian genomes. In more recent work, Dick Hwang in my lab has developed a powerful Bayesian Markov Chain Monte Carlo approach that allows systematic exploration of variation in context-dependent rates and mutational asymmetry with respect to position within an evolutionary tree and within a sequence. We have applied this to investigate variation in mutational patterns in mammalian evolution, finding in particular that CpG mutations show a reduced generation time effect relative to other mutation types (Hwang and Green, PNAS in press). I will discuss ongoing work investigating the extent to which context-dependent mutations and biased gene conversion can explain the compositional characteristics of non-functional DNA in mammals.

Our work on signals currently focuses on the computational identification of coding sequences and splicing-related motifs. We have recently begun systematic large-scale experimental testing of gene predictions (via sequencing of RT-PCR products) in selected regions of eukaryotic genomes, with the goal of determining all gene structures in these regions. The results are then used to improve our computational models. I will describe our initial work in *C. elegans*.

Protein Interactions

David Eisenberg, Peter Bowers, Michael Strong, Huiying Li, Lukasz Salwinski, Robert Riley, Richard Llwellyn, Einat Sprinzak, Todd Yeates

UCLA-DOE Institute of Genomics and Proteomics, UCLA

Protein interactions control the life and death of cells, yet we are only beginning to appreciate the nature and complexity of their networks. We have taken two approaches towards mapping these networks. The first is the synthesis of information from fully sequenced genomes into knowledge about the network of functional interactions of proteins in cells. We analyze genomes using the Rosetta Stone, Phylogenetic Profile, Gene Neighbor, Operon methods to determine a genome-wide functional linkage map. This map is more readily interpreted when clustered, revealing groups of proteins participating in a variety of pathways and complexes. Parallel pathways and clusters are also revealed, in which different sets of enzymes operate on different substrates or with different cofactors. These methods have been applied genome-wide to *Micobacterium tuberculosis* and *R. Palustris*, as well as to more than 160 other genomes. Many results are available at: http://doe-mbi.ucla.edu/pronav The outcome is increased understanding of the network of interacting proteins, and enhanced knowledge of the contextual function of proteins. The information can be applied in structural genomics to find protein partners which can be co-expressed and co-crystallized to give structures of complexes. These inferred interactions can be compared to directly measured protein interactions, collected in the Database of Interacting Proteins: http://dip.doe-mbi.ucla.edu/. These observed networks constitute a second approach to detailing protein networks.

References

Visualization and interpretation of protein networks in *Microbacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. M. Strong, T.G. Graeber, M. Beeby, M. Pellegrini, M.J. Thompson, T.O. Yeates, & D. Eisenberg (2003). *Nucleic Acids Research*, 31, 7099-7109 (2003).

Prolinks: a database of protein functional linkages derived from coevolution. P.M. Bowers, M. Pellegrini, M.J. Thompson, J. Fierro, T.O. Yeates, and D. Eisenberg (2004). *Genome Biology*, 5:R35.

In silico simulation of biological network dynamics. L. Salwinski & D. Eisenberg (2004), *Nature Biotechnology*, 22, 1017-1019.

From Cellular Networks to Molecular Interactions and Back

Hanah Margalit
The Hebrew University of Jerusalem

The recent large-scale functional genomic and proteomic experiments provide various types of genome-scale information, such as binding sites of transcription factors and protein-protein interactions. Representing these various types of data as networks of molecular relations opens new ways for exploring the mechanistic modules and the underlying evolutionary forces that shape the cellular circuitry.

To study the mechanistic modules responsible for the switching (on or off) of a variety of cellular processes we integrated the networks of protein-protein interaction and transcription regulation in *Saccharomyces cerevisiae*. Recent studies have focused on either regulatory or proteomic interactions. Yet, analyzing each of these networks separately hides the full complexity of the cellular circuitry, as many processes involve combinations of these two types of interactions. To this end we developed a new algorithm for detecting composite motifs in networks comprising two or more types of connections. Analysis of the integrated network of protein-protein interaction and transcription regulation in *S. cereviaiae* revealed several composite network motifs that may constitute the functional building blocks of various cellular processes.

Network methodology can be used to study other aspects of the cellular circuitry. By representing chromosomal adjacency of genes as a network and integrating it with the transcriptional regulatory network we revealed links between transcription regulation and chromosomal organization in both *Escherichia coli* and *S. cerevisiae*. Our findings suggest that in both organisms transcription regulation has shaped the organization of transcription units on the chromosome. Differences found between the organisms reflect the inherent differences in transcription regulation between pro- and eukaryotes.

Detailed examination of the networks provides insight into the characteristics of the molecular interactions comprising them. In turn, the knowledge gained from known interactions can be used in the development of predictive algorithms for identifying novel interactions. Application of these algorithms genome-wide will enrich the repertoire of molecular interactions and provide a more complete picture of the cellular networks. I will describe our algorithm for predicting target genes of novel transcription factors, based on their amino acid sequence and on knowledge of the binding pattern of other proteins in their family. It is possible that in the future such approaches may enable the determination of the regulatory networks in the cell based on genomic sequence data alone.

Protein-Protein Interactions: The Challenge of Predicting Specificity

Shoshana J. Wodak
Hospital for Sick Children, Toronto Ontario, Canada

Protein-protein interactions are probably amongst the most ubiquitous types of interactions and play a key role in all cellular processes. Determining the interaction network of whole organisms has therefore become a major theme of functional genomics and proteomics efforts. Computational methods for inferring protein interactions are likewise attracting much interest. Particularly remarkable has been the setup of CAPRI (Critical Assessment of PRedicted Interactions), a community-wide experiment analogous to CASP (Critical Assessment of Structure Predictions), but aimed at assessing the performance of protein-protein docking procedures. To this day seventeen complexes offered by crystallographers as targets prior to publication, have been subjected to structure prediction by docking their two components. Hundreds of predictions for these complexes were submitted by an average of 20 predictor groups and assessed by comparing their geometry to the X-ray structure and by evaluating the quality of the prediction in the regions of interaction. Over the four years of CAPRI's existence progress in the prediction quality was clearly observed, but major challenges remain. One is the ability to handle conformational flexibility, which often plays a major role. Another key challenge is to single out specific from non-specific association modes, a problem for which computational analyses are still seeking solutions. Various aspects of these challenges will be discussed and possible avenues for future progress will be outlined.

The Modern RNA World: Computational Screens for Noncoding RNA Genes

Sean R. Eddy
Washington University, St. Louis

Some genes produce RNAs that function directly as RNA rather than encoding proteins. The diversity of noncoding RNAs in nature is largely unknown, because RNA genes have been difficult to detect systematically, and most current genefinding approaches focus exclusively on protein coding genes. Genome sequence analysis, functional genomics, and new computational algorithms have enabled several recent experiments that have begun to show that RNA genes and RNA-based regulatory circuits are more prevalent that we suspected.

Towards an RNA Splicing Code

Christopher Burge
Massachusetts Institute of Technology

Most human genes are transcribed as precursors containing long introns that are removed in the process of pre-mRNA splicing. The specificity of splicing is defined in part by splice site and branch site sequences located near the 5' and 3' ends of introns. However, even considering transcripts with only very short introns, these sequences contain only about half of the information required for accurate recognition of exons and introns in human transcripts. Indeed, it is well known that human transcripts contain a vast excess of sequences that match the consensus splice site motifs as well as authentic sites yet are virtually never used in splicing – so-called 'decoy' splice sites and pairs of decoy splice sites known as 'pseudoexons'. The ability of the splicing machinery to reliably distinguish authentic exons and splice sites from a large excess of these imposters implies that sequence features outside of the canonical splice site/branch site elements must play important roles in splicing of most or all transcripts. Prime candidates for these features are exonic or intronic *cis*-elements that either enhance or silence the usage of adjacent splice sites. My lab is using a combination of computational and experimental approaches to understand the elements that control the specificity of splicing. Recently completed efforts have focused on: (i) improved modeling of the classical splice site motifs using constrained maximum entropy models[(1)]; (ii) predictive identification and SNP-based validation of exonic splicing enhancers (ESEs) in human genes[(2,3)]; and (iii) studies of variations in the sequence and organization of splicing regulatory elements between different vertebrates[(4)]. I will briefly summarize this work and related work from other labs and describe in more detail the results of a screen for exonic splicing silencers (ESSs) and the development of a first-generation RNA splicing simulation algorithm[(5)]. To systematically identify ESSs, an *in vivo* splicing reporter system was developed and used to screen a library of random decanucleotides. Screening of cells representing between one- and two-fold coverage of the ~ one million possible decanucleotides yielded 141 ESSs, 133 of which were unique. The silencer activity of over a dozen of these sequences was also confirmed in a heterologous exon context and in a second cell type. Of the unique ESS decamers, 21 pairs differed by only a single nucleotide, and most could be clustered into groups to yield seven putative ESS motifs. Some of these motifs resemble known motifs bound by the hnRNP proteins H and A1, while others appear novel. Motifs derived from the ESS decamers are enriched in pseudoexons and in alternatively spliced exons, suggesting roles in suppressing pseudoexon splicing and in regulating alternative splicing. Potential roles of ESSs in constitutive splicing were explored using an algorithm, ExonScan, which simulates splicing based on known or putative splicing-related motifs. ExonScan analysis suggests that these ESS motifs play important roles in both suppression of pseudoexons and in splice site definition. Synergistic combinations of computational and experimental approaches appear most promising for making further progress towards complete understanding of the RNA splicing code.

1. Yeo, G. and Burge, C. B. (2004). *J. Comp. Biol.* **11,** 377-394.
2. Fairbrother, W., Yeh, R.-F., Sharp, P. A. and Burge, C. B. (2002). *Science* **297,** 1007-1013.
3. Fairbrother, W. G., Holste, D., Burge, C. B. and Sharp, P. A. (2004). *PLoS Biol.* 2, e268.
4. Yeo, G., Hoon, S., Venkatesh, B. and Burge, C. B. (2004). *Proc. Natl. Acad. Sci USA* (in press).
5. Wang, Z., Rolish, M., Yeo, G., Tung, V., Mawson, M. and Burge, C. B. (2004). (unpublished data).

Discovering Evolutionary Mechanisms from Multiple Metrics of Molecular Evolution

Christopher Lee
University of California, Los Angeles

The availability of multiple genome sequences is the first essential ingredient for obtaining a detailed history of the evolutionary mechanisms that have constructed modern organisms. A second key ingredient is the development of multiple metrics of rates for different evolutionary processes, and of different types of selection pressure. We have used metrics for a wide variety of evolutionary processes--exon creation and loss; splice site movement; protein reading frame preservation; point substitution rates and selection pressures; premature termination codons, and conditional selection pressures--to examine the role of alternative splicing in the evolution of mammalian genomes. These data show that alternative splicing can produce a striking acceleration in evolution of a single exon of a gene, by reducing negative selection pressure against changes to that exon. This acceleration in the evolution of a specific protein subsequence shows clear independent evidence of adaptive benefit, that has been strongly selected for during recent evolution. Human genome data suggest that up to half of recently created exons may have been introduced through such an alternative splicing mechanism. We have also used new metrics of selection pressure to automate discovery of drug resistance mutations in HIV, and to analyze the evolutionary pathways of the viral population.

Probing The PDB

Helen M. Berman
Protein Data Bank; Research Collaboratory for Structural Bioinformatics, [a]Rutgers, the State University of New Jersey

The RCSB Protein Data Bank (PDB; www.pdb.org) is a publicly accessible information portal for researchers and students interested in structural biology. At its center is the PDB archive – the sole international repository for the 3-dimensional structure data of biological macromolecules.

This talk will focus on the tools provided by RCSB PDB to browse and explore these structures. Structures can be searched and reviewed using a variety of parameters. Data from related resources, including Gene Ontology, EC, KEGG Pathways, and NCBI are mapped to structures and loaded into the database. Structures are also linked to their corresponding entry in other databases, including Swiss-Prot, SCOP, and PubMed.

The RCSB PDB is managed by three RCSB members - Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST. Support is from the NSF, NIGMS, the Office of Science, DOE, NLM, NCI, NCRR, NIBIB, and NINDS. The RCSB PDB is a member of wwPDB.

Structural Alignment and Classification of all Known Protein Structures

Rachel Kolodny, Patrice Koehl and Michael Levitt
Stanford University

We have carried out the largest and most comprehensive comparison of protein structural alignment methods. Specifically, we evaluate six publicly available structure alignment programs: SSAP, STRUCTAL, DALI, LSQMAN, CE and SSM by aligning all 8,581,970 protein structure pairs in a test set of 2,930 protein domains specially selected from CATH v.2.4 to ensure sequence diversity. Our own method STRUCTAL has also been run on SCOP v. 1-65.

Here we use this data to discuss the importance of having an objective different geometric match measures with which to evaluate an alignment. With this improved analysis we show that there is a wide variation in the performance of different methods; the main reason for this is that it can be difficult to find a good structural alignment between two proteins even when such an alignment exists. Methods that do best in our study are neither the most popular nor those that are generally accepted to work well. We find that STRUCTAL and SSM perform best, followed by LSQMAN and CE. Our focus on the intrinsic quality of each alignment allows us to propose a new method, called 'Best-of-All', which combines the best results of all methods. Some commonly used methods miss almost half of the good 'Best-of-All' alignments.

We discuss the differences between a set of structural alignments and a classification of structures. We compare the most common classifications of protein structures (CATH, SCOP and DALI/FSSP) and show that they are really quite different. We also present preliminary results on an objective method that derives a classification of structures from a series of pair-wise alignments.

Comparative Plant Genomics: Evaluation of the Model Genome Concept

Volker Brendel
Iowa State University

The first plant genome was made available to near completion in 2000. Arabidopsis thaliana continues to be an important model system for studying genome content and organization and for functional genomics. Four years later, the rice genome is essentially finished, representing the first monocot genome and a size scale-up of threefold compared to the Arabidopsis genome. The fast expansion of the number of prokaryotic and animal genomes over a short period of time appears to have jumped over into the plant genome research field: the genome of Medicago truncatula and Lotus japonicus are also soon to be finished, sequencing projects for tomato and Physcomitrella patens have been announced, and a request for proposals is out for sequencing the maize genome (which is approximately the size of the human genome). I will discuss efforts of my group to catch up with the computational analysis of all these data. While the excitement is always with respect to the novel projects, how well do we actually understand the current genome data? How many gene models are solidly established? How large is the error rate in computational gene structure predictions? In view of inevitable transitive gene structure annotation when comparing genomes, assessments of accuracy are of paramount importance. I will discuss various tools to facilitate gene structure annotation and evaluation and present arguments for the feasibility and necessity of community-based annotation.

Lessons from the Arabidopsis Genome: Decoding Evidence for Novel Transcription

Terry Gaasterland
The Rockefeller University and Scripps Institution of Oceanography, Genome Research Center, University of California, San Diego

Several recent surveys of gene expression indicate that genome transcription activity extends well beyond mRNA, tRNA and rRNA gene expression. Large scale studies that completely tiled human chromosomes 21 and 22 onto 2-color or 1-color microarrays and hybridized with total RNA found considerable transcriptional activity in intergenic, intronic, and UTR antisense regions (Rinn et al 2003; Kapranov et al 2002). Studies that used chromatin immuno-precipitation to isolate DNA bound to selected transcription factors followed by hybridization on DNA microarrays ("ChIP-chip" studies) have found unexpected binding events in regions annotated as intergenic (Euskirchen et al 2004; Martone 2003; Kampa et al 2004; Cawley et al 2004). In contrast, ChIP-chip studies of POL-II binding sites have tended to identify primarily previously annotated coding regions (Ren & Dynlacht 2004). Other data sources include SAGE-like surveys of gene expression using the Massively Parallel Signature Sequence (MPSS) technology, which invariably yield substantial evidence for transcription outside of previously annotated genes as well as quantitative gene expression levels for annotated genes.

Some of the additional transcription is explained by the presence of small non-coding RNA genes in intergenic regions. In the case of microRNAs, transcripts from these ~150-350 nucleotide (nt) genes fold into secondary structures with long, imperfect hairpins that are processed by a protein complex that recognizes and cleaves double-stranded RNA to release short ~19-23 nt single-stranded RNA molecules. These microRNAs are complementary to mRNA transcripts and suppress protein expression by repressing translation or by triggering mRNA degradation. In plants, microRNAs tend to bind within coding regions; in animals, they bind to the 3'UTR.

This talk presents observations about control of gene and protein expression in Arabidopsis thaliana based on the following data sources: tissue specific MPSS data and Affymetrix gene expression data on stress response, genome-wide prediction of binding site clusters for known transcription factors, evaluation of alternative splicing evident in cDNA and EST sequences, and microRNA identification and mRNA target prediction (Hoth et al 2003). These data have all been combined to yield a model of microRNA regulation of gene and protein expression in plants.

1. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. The transcriptional activity of human Chromosome 22. Genes Dev. 2003 17(4):529-40.
2. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. Large-scale transcriptional activity in chromosomes 21 and 22. Science. 2002 296(5569):916-9.
3. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M. CREB Binds to Multiple Loci on Human Chromosome 22. Mol Cell Biol. 2004 24(9):3804-14.
4. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. Distribution of NF-kappaB-binding sites across human chromosome 22. Proc Natl Acad Sci U S A. 2003 100(21):12247-52.
5. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. 2004 14(3):331-42.
6. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell. 2004 116(4):499-509.
7. Ren B, Dynlacht BD. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. Methods Enzymol. 2004;376:304-15.
8. Hoth S, Ikeda Y, Morgante M, Wang X, Zuo J, Hanafey MK, Gaasterland T, Tingey SV, Chua NH. Monitoring genome-wide changes in gene expression in response to endogenous cytokinin reveals targets in Arabidopsis thaliana. FEBS Lett. 2003 554(3):373-80.

Building Genotype Phenotype Data Resources

Russ Altman
Stanford University

In the post-genome era, one of the major challenges to informatics is to support the association of genotype with phenotype. In particular, methods are required to represent and analyze data in standard formats using agreed semantics, in order to build a useful public database. Such standards may require that methods for collecting certain types of data (particularly high-throughput data) be standardized to allow for integration of data across multiple conditions. The best progress has been made in standardizing exchange of genotype and some types of phenotype data, most notably microarray expression data. However, there are significant challenges in representing other phenotype data, because the experimental methods used to collect this data are diverse, and because many biologists are not willing to constraint their scientific programs in order to be compatible with standards. This can lead to individually powerful data sets that stand alone, difficult to integrate with other data sources.

We are building the Pharmacogenomics Knowledge Base (PharmGKB, http://www.pharmgkb.org/) as an initial example of a diverse post-genome database. Pharmacogenomics is the study of how variation in the genome leads to variation in the response to drugs. The PharmGKB contains information about genotypic variation in a set of populations, and associates these with variation in phenotypes at molecular, cellular, organ and organisms levels. The PharmGKB currently contains genotyping information from more than 5000 individuals and phenotype information from more than 3000 individuals.

The goals of PharmGKB include the development of new tools for pharmacogenomics, and the mining/integration of existing databases. We have developed a method for defining haplotype tagging SNPs, and have shown that these htSNPs can be used to efficienctly recover the full genotype. We have also developed text mining algorithms to catalog all published gene-drug interactions, in order to provide comprehensive coverage of the literature in PharmGKB. The current challenges to PharmGKB include 1) defining standards for exchange of phenotype information, 2) supporting association studies for finding genotype-phenotype correlations, 3) defining and supporting the definition of drug-related pathways, 4) linking high-throughput data sources with genes, drugs and diseases of interest, and 5) linking molecular structural and cheminformatics information to pharmacogenetic variation.

Highly Expressed Genes Based on Codon Usage Biases in Archaeal and Eukaryotic Genomes

Samuel Karlin and Jan Mrázek
Stanford University

Based on rRNA sequence criteria, life has been broadly divided into the three domain: bacteria, archaea and eukaryotes, which are believed to reflect phylogenetic relationships. The archaea are further classified into Crenarchaea and Euryarchaea and recently possibly nano-archaea. For most bacterial organisms during exponential growth, ribosomal proteins (RP), transcription/translation processing factors (TF), and the major chaperone/degradation genes (CH) functioning in protein folding and trafficking tend to be highly expressed. The gene classes (RP, CH, and RF) serve as representative of highly expressed genes, and our method specifies genes with rather similar codon usages as PHX genes. These assignments are reasonable under fast growth conditions, where there is a need for many ribosomes, for proficient transcription and translation, and for many CH proteins to ensure properly folded, modified, and translocated protein products. The codon usage difference of the gene group F with respect to the gene group G is calculated by the formula $B(F|G)=\sum_a p_a(F)\left[\sum_{(x,y,z)=a}|f(x,y,z)-g(x,y,z)|\right]$, where $\{p_a(F)\}$ are the average amino acid frequencies of the genes of *F*. Predicted expression levels with respect to individual standards can be based on the ratios $E_{RP}(g)=B(g|C)/B(g|RP)$, $E_{CH}(g)=B(g|C)/B(g|CH)$, and $E_{TF}(g)=B(g|C)/B(g|TF)$, where *C* is the totality of all genes of the genome. Using these gene classes as standards, a gene is Predicted Highly eXpressed (PHX) if its codon usage is rather similar to at least two of the RP, TF, and CH gene classes and deviates strongly from the average gene of the genome. An overall estimate of the expression level of the gene *g* is $E(g)$ defined by the equation $B(g|C)/E(g)=(1/3)(B(g|RP)+B(g|TF)+B(g|CH))$.

The criterion $E(g)>1$ and where at least two of the values $E_{RP}(g)$, $E_{TF}(g)$, $E_{CH}(g)$ exceed 1.05 provides an excellent benchmark in reflecting high protein molar abundance in a rapid growth environment.

In all currently available archaeal genomes, the thermosome chaperonin genes rank among the top PHX. DnaK (HSP70) is found PHX virtually only in archaeal mesophiles or in archaeal moderate thermophiles. The Lon protease is absent from the Crenarchaea but usually PHX among the euryarchaea. Archaea genomes are also pervasive with proteasome units. Other distinctive proteins of archaea generally PHX and absent from bacteria highlight PCNA (proliferating cell nuclear complex) a replication auxiliary factor (sliding clamp subunit) responsible for tethering the catalytic subunit of DNA polymer to DNA during high-speed replication. The ribosomal protein P_0 (acidic, regulatory) whereas the ribosomal machinery in eukaryotes contains P_0, P_1, P_2 featuring a hyperacidic run at its carboxyl end. Other distinctive PHX genes found in all archaea: Cdc48 Cell division control protein 48; Cdc6 Replication initiation; RadA DNA repair and recombination protein in archaea.

In prokaryotes, the maximum $E(g)$ level correlates negatively with the doubling time of the organisms. Compared to bacterial genomes, relatively few RP genes of archaea are PHX and many are expressed as average genes. A clear exception is *M. maripaludis*. The yeast genome parallels *E. coli* in PHX genes plus the addition of actin, cofilin and related genes. The most PHX genes of *Drosophila* encode the cytoskeletal proteins.

FRONTIERS OF BIOINFORMATICS: UNSOLVED PROBLEMS AND CHALLENGES

October 15-17, 2004

Poster Session

1. Chen, Lamei (University of California, Los Angeles); *Analyze HIV-1 Mutation Evolution as a Conditional Selection Pressure Network.*

2. Chuang, Jeffrey (University of California, San Francisco); *Mutation Rates are Correlated in Mammalian Lineages.*

3. Day, Ryan (University of Washington); *Molecular Dynameomics*.

4. Kechris, Katherina (University of California, San Francisco); *Analysis of human alternative splices predicted from exon junction arrays.*

5. Liang, Mike (Stanford University); *Integrating Sequence and Structure Data for Annotating Functional Sites on Protein Structures.*

6. Liu, Shuo (Stanford University); *Processes and Functions Potentially Regulated by Alternative Splicing Uncovered Through Study of Protein Domains.*

7. Lotan, Itay (University of California, Berkeley); *Real Space Protein Model Completion: an Inverse Kinematics Approach.*

8. Mrázek, Jan (Stanford University); *Genomic Comparisons among γ-proteobacteria.*

9. Naughton, Brian; *MotifCut: Motif Finding and Spectral Graph Theory.*

10. Reyes, Vicente M. (University of California, San Diego); *Whole Proteome Functional Annotation via Automated Detection of Ligand 3-D Binding Site Motifs: Application to ATP- and GTP-Binding Sites in Unannotated Proteins of Dictyostelium discoideum.*

11. Saxonov, Serge (Stanford University); *SampleScan: A Sampling Approach to Motif Discovery in Nucleotide Sequences.*

12. Veretnik, Stella (University of California, San Diego); *Assignment of structural domains in proteins: why is it so difficult.*

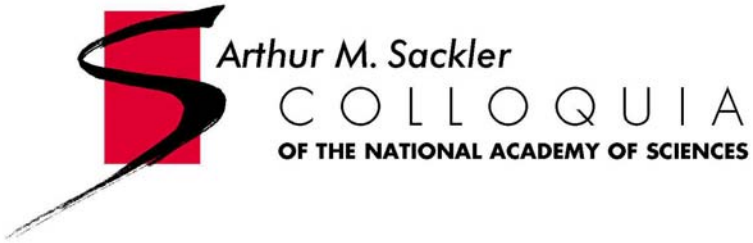
13. Wang, Qi (University of California, Los Angeles); *Detecting Tissue-Specific Regulation of Alternative Splicing as a Qualitative Change in Microarray Data.*

14. Xing, Yi (University of California, Los Angeles); *alternative splicing opens neutral paths for genome evolution.*

15. Xu, Na (University of California, Berkeley); *Identifying functional importance of NCS conserved across multiple species.*

16. Yeh, Iwei (Stanford University); *A Cellular Architecture Ontology for Analyzing Protein-Protein Interactions Based on Subcellular Localizations.*

17. Zhao, Keyan (University of Southern California); *Genome-wide association mapping of flowering time in model plant - Arabidopsis thaliana.*



FRONTIERS OF BIOINFORMATICS: UNSOLVED PROBLEMS AND CHALLENGES

October 15-17, 2004

POSTER

ABSTRACTS

This page is intentionally left blank.

Analyze HIV-1 Mutation Evolution as a Conditional Selection Pressure Network

L. Chen, C.J. Lee
Department of Chemistry & Biochemistry, University of California, Los Angeles

Antiretroviral therapy of HIV-1 frequently results in the emergence of drug resistant variants from the viral quasispecies. The development and maintenance of drug resistance usually requires the accumulation of 2 or more mutations. Many drug resistant mutation patterns have been reported, but this information is limited and static. It is useful to obtain a global picture of all the ways the viral population could respond to the current drug treatment. To address this problem, we have developed a conditional selection pressure (K_a/K_s) approach that measures how mutation at one site alters the selection pressure at another site. The conditional K_a/K_s analysis shows that different evolutionary paths to the same final genotype can have very different rates, so some paths are preferred whereas the others are not favored. We have generated a directed diagram to represent the mutation network of HIV-1 protease based on the conditional K_a/K_s analysis. This diagram shows the speeds of all possible paths of evolution the viral population will follow under the pressure of current drug treatment. This evolution network also reveals *kinetic traps*, individual sites (or groups of sites) that accumulate mutations rapidly, but which lack fast paths to drug resistance mutations (i.e. the rate constants are much slower than from wildtype). We can combine the kinetic map with existing information about specific drug resistance relationships and this may reveal general strategies for slowing the evolution of drug resistance.

Molecular Dynameomics

Ryan Day, David A. C. Beck, Stephen Edwards, Daigo Inoyama, R. Dustin Schaeffer, Robert E. Steward, George W. N. White, and Valerie Daggett

Molecular dynamics simulations allow effective characterization of the dynamics of proteins in water. They have been useful both in characterizing the native state dynamics of proteins and the unfolding process. Simulations have also led to models of the partially unfolded disease states of certain proteins. Molecular dynamics simulations have generally only been considered in the context of the particular system under study, however, limiting their applicability to broader questions of protein folding and dynamics. We have begun an effort to simulate a large number of proteins with different topologies under native and denaturing conditions in order to address this shortcoming. These simulations will be analyzed in a broader context in order to determine general dynamic properties of amino acids and proteins in water. We are calling this effort molecular dynameomics. The thirty most populated folds in the PDB represent 50% of the structures deposited. We have begun our work with simulations of thirty target proteins from these thirty folds. Here we present preliminary results from this set.

Analysis of human alternative splices predicted from exon junction arrays

Katherina Kechris, Jean Yee Hwa Yang, Ru-Fang Yeh
University of California, San Francisco

Following transcription, alternative splicing of exons in a pre-mRNA transcript can create multiple different protein products. This process is an important mechanism which contributes to the protein complexity found in humans. Sequence elements in exons and adjacent introns are critical for regulating alternative splicing. To discover these elements genome-wide, we use the experimental results from Rosetta's exon-junction array (Johnson et al., 2003) to first identify alternatively and constitutively spliced exons. By applying a variation of their linear model to the data, we specify a score that measures the occurrence of alternative splicing in each gene. Based on a ranking using this score, we identify genes, with their corresponding exons, that are either alternatively or constitutively spliced. Then, by comparing the word counts between these two sets of exons, and their neighboring introns, we find motifs that are associated with alternative splicing and are potential regulators. In particular, we find that adjacent introns of alternatively spliced exons are A/T rich, while those from constitutively spliced exons are G/C rich. For alternatively spliced exons, the motifs discovered tend to be A/G rich and are similar to known exonic sequence enhancers that are naturally occurring or discovered by SELEX.

Integrating Sequence and Structure Data for Annotating Functional Sites on Protein Structures

Mike Liang
Stanford University

Structural genomics initiatives are developing high-throughput methods for large-scale determination of all protein structures. The biological roles for many of these proteins are still unknown, and high-throughput computational methods for determining their function are necessary. Understanding the function of these proteins will have profound impact in drug development and protein engineering.

Current methods for functional annotation of these protein structures are based on sequence only or structure only analysis. Although sequence based methods have been quite powerful, they often have limited use when sequence similarity is low. Structure base methods are less sensitive to sequence similarity, but many of the structure based methods require manual creation of models and are thus limited by the number of available functional models. Methods for function annotation at a structural-genomics scale will require both greater sensitivity than what sequence only methods provide and more functional models than what the structural models offer.

To address the requirements for structural-genomics scale function annotation, we have developed a method, SeqFEATURE, for automatically constructing a three-dimensional (3D) model of the functional site by integrating sequence and structure data. The 3D models describe the physicochemical environment around sequence motifs and identify the significant properties that are statistically conserved or absent in the functional site. These 3D models have better sensitivity than one-dimensional (1D) sequence motifs in function annotation. By automatically creating these 3D models from sequence motifs, we have developed a method for building a library of models that can be used in context of a structural genomics pipeline for functional annotation of protein structures.

SeqFeature is available on the web at http://feature.stanford.edu/webfeature. Biologists can rapidly annotate their structure with the currently available library of functional models.

Processes and Functions Potentially Regulated by Alternative Splicing Uncovered Through Study of Protein Domains

Shuo Liu
Stanford University

Alternative splicing plays an important role in processes such as development, differentiation, and cancer. With the recent increase of the estimates of the number of human genes that undergoes alternative splicing from 5% to 74%, it becomes critical to develop a better understanding of its functional consequences and regulatory mechanisms. We conducted a large scale study of the distribution of protein domains in a curated data set of several thousand genes and identified protein domains disproportionately distributed among alternatively spliced genes. We also identified a number of protein domains that tend to be spliced-out. Both the proteins having the disproportionately distributed domains as well as those with spliced-out domains are predominantly involved in the processes of cell communication, signaling, development and apoptosis. These proteins function mostly as enzymes, signal transducers, and receptors.

Real Space Protein Model Completion: an Inverse Kinematics Approach

Henry van den Bedem Itay Lotan†, Jean-Claude Latombe† and Ashley Deacon**
** Joint Center for Structural Genomics, Stanford Synchrotron Radiation Laboratory, SLAC, † Department of Computer Science, Stanford University*

Rapid protein structure determination relies greatly on software that can automatically build a protein model into an experimental electron density map. In favorable circumstances, various software systems are capable of building over 90% of the final model. However, completeness falls off rapidly with the resolution of the diffraction data. Manual completion of these partial models is usually feasible, put is time-consuming, and prone to subjective interpretation. Except for the N- and C-termini of the chain, the end points of each missing fragment are known from the initial model. Hence, fitting fragments reduces to an inverse kinematics problem.

We have combined fast, inverse kinematics algorithms with a real space, torsion angle refinement procedure in a two stage approach to fit missing main-chain fragments into the electron density between two anchor points. The first stage samples a large number of closing conformations, guided by the electron density. These candidates are ranked according to density fit. In a subsequent refinement stage, optimization steps are projected onto a carefully chosen subspace of conformation space to preserve rigid geometry and closure.

In a test set of 103 structurally diverse fragments within one protein, the algorithm closed gaps of 12 residues in length to within, on average, 0.52Å all-atom Root Mean Square Deviation (aaRMSD) from the final, refined structure at a resolution of 2.8Å. The algorithm has also been tested and used to aid protein model completion in areas of weak or ambiguous experimental electron density, where an initial model was built using ARP/wARP (Perrakis et al., 1999) or RESOLVE (Terwilliger, 2002). At a resolution of 2.4Å, it closed a 10-residue gap to within 0.43Å aaRMSD of the final, refined structure. In another case, a 14-residue gap in a 51%-complete model built at 2.6Å was closed to within 0.9Å aaRMSD. Our method was furthermore used to correctly identify and build multiple, alternative main-chain conformations at a resolution of 1.8Å

Genomic Comparisons among γ-proteobacteria

Jan Mrázek[1], Alfred M. Spormann[2] and Samuel Karlin[1]

[1]Department of Mathematics and [2]Departments of Civil and Environmental Engineering, of Biological Sciences, and of Geological and Environmental Sciences, Stanford University

Highly expressed genes in most unicellular and some multicellular organisms exhibit characteristic codon usage biases that distinguish them from the bulk of genes in a genome. We have developed a method to identify predicted highly expressed (PHX) genes in complete genomes. Predicted highly expressed (PHX) genes are compared for sixteen γ-proteobacteria and their similarities and differences are interpreted with respect to known or predicted physiological characteristics of the organisms. PHX genes often reflect the organism's lifestyle, habitat, nutrition sources and metabolic propensities. This technique allows to predict predominant metabolic activities of the microorganisms operating in their natural habitats. Among the most striking findings is an unusually high number of PHX enzymes acting in cell wall biosynthesis, amino acid biosynthesis and replication in the ant endosymbiont *Blochmannia floridanus*. We ascribe the abundance of these PHX genes to specific aspects of the relationship between the bacterium and its host. *Xanthomonas campestris* is also unique with very high number of PHX genes acting in flagellum biosynthesis, which may play a special role during its pathogenicity. *Shewanella oneidensis* possesses three protein complexes which all can function as complex I in the respiratory chain but only the Na^+-transporting NADH:ubiquinone oxidoreductase *nqr*-2 operon is PHX. The PHX genes of *Vibrio parahaemolyticus* are consistent with the microorganism's adaptation for extremely fast growth rates. Comparative analysis of PHX genes from complex environmental genomic sequences as well as from uncultured pathogenic microbes can provide a novel, useful tool to predict global flux of matter and key intermediates, as well as specific targets of antimicrobial agents.

MotifCut: Motif Finding and Spectral Graph Theory

Brian Naughton

Computationally identifying conserved motifs in DNA sequences is important for understanding gene regulation. For example, it is used to find conserved DNA motifs in the upstream sequences of co-expressed genes from microarray and chromatin immunoprecipitation (CHIP) experiments. We have approached the motif-finding problem in a novel way, taking inspiration from an image analysis algorithm used to identify the foreground of an image and separate it from the background. We build a graph where all of the words k bases long (k-mers) in the sequence are represented as nodes. K-mers that are similar to each other are then connected with an edge in the graph. Usually, motif finding algorithms search the DNA sequence for a set of k-mers that are as similar to each other as possible. More advanced algorithms may also ensure that the motif is different from a model of the background sequences. We identify groups of k-mers that are similar to each other (the "foreground") but also as different as possible from all of the other k-mers in the graph (the "background"). This problem is computationally difficult (NP complete), so we use an eigendecomposition as an approximation to the exact solution. MotifCut shows improved results over other motif-finding methods under many conditions.

Whole Proteome Functional Annotation via Automated Detection of Ligand 3-D Binding Site Motifs: Application to ATP- and GTP-Binding Sites in Unannotated Proteins of *Dictyostelium discoideum*

Vicente M. Reyes
University of California, San Diego

The first few years of the new millennium have seen the avalanche of genome sequence data, a trend that is predicted to continue for the next 10 to 15 years. Assignment of function ("annotation") to these mostly novel sequences is fast becoming top priority, for obvious basic and medical reasons. However, due to the sheer number of sequences to be annotated, conventional functional assay techniques involving gene cloning, protein expression and purification, etc., become extremely impractical even if implemented in high-throughput fashion including robotics. The solution lies in tapping the ever-increasing power of computers and the versatility of "smart" databases to predict the function of novel sequences, especially proteins, on the basis of their primary sequences alone.

The present study, a part of the grand-challenge initiative "The Encyclopedia of Life" at the San Diego Supercomputer Center, nicely illustrates whole proteome functional annotation using computational techniques. Unlike DNA, which generally function at the level of primary sequence, and RNA, which (generally) function at the level of secondary structure, proteins function at the level of tertiary structure. The first step in our study is therefore the prediction of the 3-D structure of all the ORFs in the genome of our test organism, *Dictyostelium discoideum*, via a combination of homology modeling and threading algorithms, and a final step of model completion using the program Modeller6v2. These predicted 3-D structures will then be screened in a later step for various ligand 3-D binding site motifs. The second step in our study is therefore the construction of the 3-D binding site motif of a given ligand using as "training set" several (typically 8-12, but see below) experimentally solved protein structures with the ligand of interest bound.

As mentioned, the third and final step in our study is the screening of the predicted 3-D structures of the *D. discoideum* proteins (from step 1) for the 3-D binding site motif (from step 2) using a set of Fortran 77 and 90 programs we developed. The program set treats the motif as a tree structure with a root, nodes, branches and edges, and searches for such “tree” in the protein structures. Since the latter are predicted and as such carry significant degrees of uncertainty, we have tried to minimize the effects of such uncertainties by incorporating fuzzy logic into the screening process. This was done by using two types of reduced representation of the proteins in the test set, namely: (a.) representing the protein as a collection of the centroids of its constituent amino acids, and (b.) representing the protein as the aggregate of its backbone atoms and the centroids of the side chains of its constituent amino acids. Another quality assurance step is the use of the program FADE, which detects crevice and pocket residues in protein 3-D structures. In building the binding site motif, we make sure that at least one (preferably two or more) residues making it up is/are located in crevices/pockets in an effort to reduce false positives, the rationale being that binding site residues are almost always found in such locations.

We chose ATP and GTP as the pilot ligands for this study. As the ATP-binding protein family is quite heterogeneous, we first narrowed down our study to (1) ser/thr protein kinases (PKs), (2) cAMP-dependent PKs, and (3) ABC transporters. The training set for (1) contained sufficient members, but those of (2) and (3) contained only one member each, as there was only one experimentally solved structure each of a cAMP-dependent PK and an ABC transporter, with bound ATP, on deposit in the PDB. Nevertheless, binding site motifs for all three families were successfully built.

Control screens were then performed. Positive controls composed of ser/thr PKs of available experimental 3-D structures, and negative controls composed of proteins known not to bind ATP, also with available experimental 3-D structures, were subjected to the screening procedure. Both yielded the expected results, albeit with a small proportion of false negatives, and an even smaller proportion of false positives.

The 400 previously unknown and unannotated proteins of *D. discoideum* were then chosen as the pilot test set. Two binding site motifs were deduced from the ser/thr PK training set, one with the two ribose hydroxyls "bound" to a single protein residue centroid (type 1), and a second where the two hydroxyls are separately "bound" (type 2). Screening the test set for the ser/thr ATP binding site motif type 1, our program detected 32 putative ATP-binding proteins, which on closer human inspection, revealed that only 15 are true positives. Screening for type 2, the program picked up 25 putatives, which on closer inspection revealed that all but 1 may be false positives. It now remains to be seen whether the 16 true positives (15 type 1 and 1 type 2) picked up by our program above indeed bind ATP via actual laboratory experimental techniques.

Interestingly, our program did not detect any putative cAMP-dependent PK nor ABC transporter from the test set. This may be due to the fact that the training sets for these two families each contained only one member (as those were the only ones available from the PDB), making the screen overly specific and therefore quite insensitive.

Similar work on GTP is underway. We are also currently trying to incorporate all the different programs for all the different steps in the entire procedure into one main calling script in order to streamline the screening procedure and make it less dependent on human intervention, and therefore more amenable to complete automation, and, in turn, more suited to large-scale, whole-proteome screening.

SampleScan: A Sampling Approach to Motif Discovery in Nucleotide Sequences

Serge Saxonov, Serafim Batzoglou, Douglas L Brutlag
Stanford University

When looking for DNA motifs on a genomic scale one is faced with two main challenges. One is that most motifs are likely to be present in only a small number of sequences, making it harder for conventional motif finders to spot them. The second is that the sheer size of the data often precludes the application of many well-established algorithms. In this work we present a sampling-based method geared specifically toward discovery of motifs in large sequence sets.

In an outline, the method works by constructing many small subsets of sequences by sampling from the whole set, followed by an application of a motif-finder to each of the subsets. The motivation behind the approach is that a motif that is present in too small a fraction of the sequences to be discovered by a conventional motif-finder, will be enriched in some of the subsets, allowing it to be discovered. Each of the candidate motifs is tested and refined by scanning the entire sequence set. To help with this approach we constructed a new motif finder that outperforms other programs when run on small data sets.

As a test case we have applied the SampleScan approach to the set of yeast upstream regions. We have shown that the method can recover a substantial fraction of known yeast motifs. In addition, we have used comparative genomics information and location bias to validate the motifs.

Assignment of structural domains in proteins: why is it so difficult?

Stella Veretnik[1], Ilya N. Shindyalov[1]. Nickolai N. Alexandrov and Phillip E. Bourne, [1,2]

[1] San Diego Supercomputer Center, University of California, San Diego, [2] Department of Pharmacology, University of California, San Diego

Structural domains are often considered to be basic units of protein structure. Assignment of structural domains from atomic coordinates is crucial for understanding protein evolution and function. Currently there is no good agreement among different assignment methods for what constitute the basic structural unit, underscoring the complexity of structural domain assignment. This work discusses tendencies of individual methods and highlights the problematic areas in assignment of structural domains by experts as well as by fully automated methods.

Domain assignments were analyzed for three automatic methods (DALI[1], DomainParser[2], PDP[3]) and three expert methods (AUTHORS[4], CATH[5], SCOP[6]), using a 467-chains dataset assigned by all 6 methods. The following features were investigated: agreement on the number of assigned domains, agreement on domain boundaries, distribution of domain sizes and tendency toward assignment of discontinuous domains. Consensuses among automatic, expert and all methods were defined and used during comparison to tease out the behaviors specific to individual assignment methods or groups of methods.

We observe that unambiguous domain assignments (when all methods agree on domain assignment) are confined predominantly to one-domain chains. Agreements among all methods in multi-domain chains are infrequent; in all cases the domains are compact and clearly spatially separated. For the majority of multi-domain proteins, there is no agreement on domain assignment among all methods. From the consensus analysis we observe that the majority of the difficulties of fully automated methods stem from overwhelming reliance on the structural cues (compactness/contact density) during domain assignments and the lack of functional/evolutionary information. Thus the cases in which domains are positioned close together are difficult or impossible for automatic methods to resolve. On the other hand, the differences in expert methods arise from different philosophical approaches underlying the specific methods. Authors of the structures (AUTHORS method) tend to define domains based on functionality, which may produce small and structurally not clearly defined domains. The creators of SCOP, on the other hand, often look for the largest common structure (fold) as a domain, which often consists of several distinctive structural units. The CATH method appears to strike a balance between sometimes contradictory structural, functional and evolutionary information. The inconsistencies in expert assignments are well reflected in the propensities of different fully automated methods, as those are trained and validated using a specific expert method, thus reflecting its philosophical biases. Detailed analysis of structures which do not have consensus between the assignment methods regarding the number of assigned domains indicates the following problematic areas: (1) assignment of small domains, (2) discontinuous domains and unassigned regions in the structure, (3) splitting of the secondary structure elements between domains (if required), (4) convoluted domain interfaces and complicated architectures. Comprehensive domain re-definition, which takes into account the above issues is overdue and will be a great step toward improvement of domain definitions in multi-domain proteins, which represent (by an estimation [7]) 66-75% of the sequence database. Also, the intensive growth of 3D protein data demands fully automated approaches to be used to maintain currency and uniformity of domain information relative to the PDB.

REFERENCES

[1] Holm L., S. C. 1996 Mapping the protein universe. *Science* 273, 595-602.

[2] Guo, J-T. Xu, D. Kim, D. Xu, Y. 2003 Improving the Performance of DomainParser for Structural Domain Partition Using Neural Network, *Nucleic Acids Res*.31(3), 944-952.

[3]. Alexandrov, N. & Shindyalov, I. 2003 .PDP: protein domain parser. *Bioinformatics* 19, 429-430.

[4] Islam, S. A., Luo, J. & Sternberg, M. J. 1995 Identification and analysis of domains in proteins. *Protein Eng* 8, 513-25.

[5] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. 1997 CATH—a hierarchic classification of protein domain structures. *Structure* 5, 1093-108.

[6] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-40.

[7] Chothia C, Gough J, Vogel C, Teichmann SA. 2003 Evolution of the protein reprertoire. Science 300, 1701-03.

Detecting Tissue-Specific Regulation of Alternative Splicing as a Qualitative Change in Microarray Data

Qi Wang
University of California, Los Angeles

Alternative splicing has recently emerged as a major mechanism of regulation in the human genome, occurring in perhaps 40-60% of human genes. Thus microarray studies of functional regulation should in principle be extended to detect not only changes in the overall expression of a gene, but changes in its splicing pattern between different tissues. However, since changes in the total expression of a gene and changes in its alternative splicing can be mixed in complex ways among a set of samples, separating these effects can be difficult, and is essential for their accurate assessment. We present a simple and general approach for distinguishing changes in alternative splicing from changes inexpression, based on detecting systematic anti-correlation between two different samples' log ratios versus a pool containing both samples. We have tested this analysis method on microarray data for five human tissues, generated using a standard microarray platform and experimental protocols previously shown to be sensitive to alternative splicing. Our automatic analysis was able to detect a wide variety of tissue-specific alternative splicing events such as exon skipping, mutually exclusive exons, alternative 3' and alternative 5'splicing, alternative initiation and alternative termination, all of which were validated by independent reverse-transcriptase PCR experiments, with validation rates of 70 - 85%.Our analysis method also enables hierarchical clustering of genes and samples by the level of similarity of their alternative splicing patterns, revealing patterns of tissue-specific regulation that are distinct from those obtained by hierarchical clustering of gene expression from the same microarray data.

Alternative Splicing Opens Neutral Paths for Genome Evolution

Yi Xing
University of California, Los Angeles

The role of alternative splicing in evolution has become a subject for many recent investigations. It has been proposed that alternative splicing can reduce negative selection pressure within selected portions of a gene, accelerating its rate of evolution. Here we discuss several lines of evidence that support this hypothesis. Alternative splicing is frequently associated with recent creation and loss of exons in mammals. Alternative splicing relieves negative selection pressure against premature protein truncations, to the extent similar to that produced by diplody. Alternatively spliced regions undergo relaxed purifying selection pressure compared to other portions of the gene. Our data suggest that alternative splicing is able to open neutral paths for evolution, a principle that can add significantly to current theory of molecular evolution.

Identifying functional importance of NCS conserved across multiple species

Na Xu
University of California, Berkeley

One of the goals in computational biology is to identify the functional elements in genomes. Highly conserved non-coding sequences (NCS) across multiple species are good candidates for functional regulatory regions. We examined 2094 NCS conserved among human, mouse and Fugu, explored their features, and attempted to pick out true functional regions.

We gathered Gene Ontology (GO) and gene expression data for the neighbor genes of the NCS. With high statistical significance, we found that NCS neighbor genes are over-represented in several GO categories: transcription regulator activity, development and binding. Further analysis using expression data showed that the NCS genes are more likely to be over-expressed in nerve and brain tissues.

We related the two analyses and combined these with other factors such as NCS density. Simple significance tests and clustering methods have provided us with an initial filtering of NCS based on the functional annotation and experimental information. These results provide a promising first set of NCS examples for further exploration.

A Cellular Architecture Ontology for Analyzing Protein-Protein Interactions Based on Subcellular Localizations

Iwei Yeh and Russ B. Altman
Stanford University

We introduce a cellular architecture ontology that encodes knowledge about cellular components, including membranes, spaces, and membrane-bound compartments, and their spatial relationships to each other. This ontology facilitates computational reasoning based on protein localization on a large scale and in a systematic fashion. To demonstrate the usefulness of our ontology in automatic reasoning, we developed rules to define the accessibility of cellular components and to define the accessibility and location of proteins. Using these rules, we automatically evaluated the physical accessibility of *Saccharomyces cerevisiae* proteins based on localizations provided by Saccharomyces Genome Database (SGD) and the accuracy of protein-protein interactions from the Database of Interacting Proteins (DIP). We found areas of inconsistency between these data resources and proposed refinement of protein localizations based on localizations of interacting proteins. In some cases our ontology allowed us to propose novel localizations using simple logical rules. Our cellular architecture ontology contains links to the Gene Ontology (GO), but provides a much richer framework for supporting computational inference.

Genome-wide association mapping of flowering time in model plant - Arabidopsis thaliana

Keyan Zhao[1], María-José Aranzana[1], Sung Kim[1], John Molitor[2], Paul Marjoram[2], Fengzhu Sun[1], Magnus Nordborg[1]

[1]Molecular and Computational Biology Program; [2]Department Of Preventive Medicine, University of Southern California, Los Angeles

A genome-wide association mapping study was conducted to search for genes controlling flowering time in Arabidopsis thaliana. We have applied several algorithms based on haplotype sharing. Using 906 fragments of sequenced polymorphism data from 95 accessions, we found some strong peaks in genes known to control flowering time in A. thaliania. Increasing the sample size to 192 still gave strong signals in genes FRI and other genes. The clustering algorithms successfully detected 2 known functional alleles in FRI gene. We have also found some other interesting regions associated with flowering time by our algorithm, although they may need further validation through experimental crosses. Population structure remains a challenging issue producing spurious associations between genotype and phenotype. This study demonstrates the promise of using LD mapping to study the genetic basis of complex traits and potentially genome-wide disease association.

FRONTIERS OF BIOINFORMATICS: UNSOLVED PROBLEMS AND CHALLENGES

Participant Roster

Bruce Alberts
President
National Academy of Sciences
500 Fifth Street NW
Washington, DC 20001
Phone: (202) 334-2100
Fax: (202) 334-1647
E-mail: balberts@nas.edu

Russ Altman
Assistant Professor of Medicine
Stanford University
Section on Medical Informatics
251 Campus Drive
MSOB X-215
Stanford, CA 94305
Phone: (650) 725-3394
Fax: (650) 725-7944
E-mail: russ.altman@stanford.edu

Chitta Baral
Professor
Arizona State University
Department of Computer Science and Engineering
699 S. Mill Avenue
Tempe, AZ 85281
Phone: (480) 727-6047
Fax: (480) 965-2751
E-mail: chitta@asu.edu

Helen Berman
Professor
Rutgers University
Department of Chemistry
Wright-Reiman Labs
610 Taylor Road
Piscataway, NJ 08854
Phone: (732) 445-4667
Fax: (732) 445-4320
E-mail: berman@rcsb.rutgers.edu

Peter Bickel
Professor of Statistics
University of California, Berkeley
Evans Hall
Berkeley, CA 94720
Phone: (510) 642-1381
Fax: (510) 642-7892
E-mail: bickel@stat.berkeley.edu

Peer Bork
European Molecular Biology Laboratory
Meyerhofstrasse 1
Heidelberg
Germany
Phone: 49 [0] 6221 3870
Fax: 49 [0] 6221 3878306
E-mail: bork@embl-heidelberg.de

Serdar Bozdag
University of California, Riverside
3131 Watkins Drive
Apartment 231
Riverside, CA 92507
Phone: (951) 276-0470
E-mail: sbozdag@cs.ucr.edu

Volker Brendel
Bergdahl Professor of Bioinformatics
Iowa State University
Department
of Genetics, Development and Cell Biology
and Department of Statistics
2112 Molecular Biology Building
Ames, IA 50011-3260
Phone: (515) 294-9884
Fax: (515) 294-6755
E-mail: vbrendel@iastate.edu

Roy Britten
Professor Emeritus
California Institute of Technology
101 Dahlia Avenue
Corona del Mar, CA 92625
Phone: (949) 675-2159
Fax: (949) 675-1837
E-mail: r.britten@comcast.net

Christopher Burge
Assistant Professor of Biology
Massachusetts Institute of Technology
Room 68-223A
Cambridge, MA 02138
Phone: 617-258-5997
E-mail: cburge@mit.edu

Huangming Chen
Research Assistant III
The Salk Institute
10010 North Torrey Pines Rd
La Jolla, CA 92037
Phone: (858) 453-4100
E-mail: hchen@salk.edu

Lamei Chen
University of California, Los Angeles
MBI #609
Los Angeles, CA 90095
Phone: (310) 794-4206
Fax: (310) 267-0248
E-mail: lchen@chem.ucla.edu

Jeffrey Chuang
Postdoctoral Fellow
University of California, San Francisco
Department of Biochemistry and Biophysics
600 16th Street
San Francisco, CA 94143-2240
Phone: (415) 514-2616
Fax: (415) 514-2617
E-mail: jchuang@genome.ucsf.edu

Valerie Daggett
Professor of Medicinal Chemistry
University of Washington
H165B, Box 357610
Seattle, WA 98195-7610
Phone: (206) 685-7420
Fax: 206) 685-3252
E-mail: daggett@u.washington.edu

Ryan Day
University of Washington, Seattle
Biomolecular Structure and Design Program
Department of Medicinal Chemistry
Seattle, WA 98195-7610
E-mail: rd@u.washington.edu

Russell F. Doolittle
Research Professor
University of California, San Diego
Department of Chemisry and Biochemistry
9500 Gilman Drive
2040 Urey Hall Addition
La Jolla, CA 92093
Phone: (858) 534-3575
Fax: (858) 534-6255
E-mail: rd@zeus.ucsd.edu

Debojyoti Dutta
University of Southern California
1042 W. 36th Place
Los Angeles, CA 90089
Phone: (213) 248-3656
Fax: (213) 740-2437
E-mail: ddutaa@usc.edu

Sean Eddy
Professor of Bioinformatics
Washington University School of Medicine
Department of Genetics
4566 Scott Avenue
St. Louis, MO 63110
Phone: (314) 362-7666
E-mail: eddy@genetics.wustl.edu

David Eisenberg
Professor of Biological Chemistry and
Molecular
Biology and Director
UCLA-DOE Laboratory of Structural Biology
and Molecular Medicine
University of California, Los Angeles
Box 951570
Los Angeles, CA 90095
Phone: (310) 825-3754
Fax: (310) 206-3914
E-mail: david@mbi.ucla.edu

Iddo Friedberg
The Burnham Institute
10901 North Torrey Pines Road
La Jolla, CA 92037
Phone: (858) 646 3100 x3516
Fax: (858) 713 9930
E-mail: idoerg@burnham.org

Terry Gaasterland
Assistant Professor and Head
Laboratory of Computational Geonomics
The Rockefeller University
1230 York Avenue
New York, NY 10021
Phone: (212) 327-7755
Fax: (212) 327-7765
E-mail: gaasterland@rockefeller.edu

Mark Gerstein
Principal Investigator
Yale University
Molecular Biophysics and Biochemistry
Bass 432A, 266 Whitney Ave.
New Haven, CT 06520
Phone: (203) 432 6105
Fax: (360) 838 7861
E-mail: mark.gerstein@yale.edu

Margaret Goodman
Associate Professor of Biology
Wittenberg University
Box 720
Springfield, OH 45501
Phone: (937) 327-6142
Fax: (937) 327-7522
E-mail: mgoodman@wittenberg.edu

Phillip Green
Investigator, Howard Hughes Medical Institute
University of Washington Medical School
Department of Molecular Biotechnology
Box 357730 (HSB K343B)
1959 Pacific Street, N.E.
Seattle, WA 98195
Phone: (206) 685-4341
Fax: (206) 685-9720
E-mail: phg@u.washington.edu

David Haussler
Howard Hughes Medical Institute
Investigator. Director, Center for Biomolecular Science & Engineering.
Professor, Computer Science
Univesrity of California, Santa Cruz
Center for Biomolecular Science and Engineering
321 Baskin Engineering Bldg
Santa Cruz, CA 95064
Phone: (831) 459-2105
Fax: (831) 459-4829
E-mail: haussler@cse.ucsc.edu

Haiyan Hu
University of Southern California
1042 W. 36th Place, DBR 202
Los Angeles, CA 90007
Phone: (213) 399-6627
E-mail: hhu@usc.edu

Yu Huang
University of Southern California
1042 W. 36th Place, DBR 202
Los Angeles, CA 90089
Phone: (213) 821-3167
E-mail: yuhuang@usc.edu

Steve Jacobsen
University of California, Los Angeles
Department of Molecular, Cellular and Developmental Biology
P. O. Box 951606
Los Angeles, CA 90095
Phone: (310) 825-0182
E-mail: jacobsen@ucla.edu

Rui Jiang
University of Southern California
1042 W. 36th Place, DBR 289
Los Angeles, CA 90089
Phone: (213) 821-2229
Fax: (213) 740-2437
E-mail: ruijiang@usc.edu

Jason Johnson
Scientific Director
Rosetta Inpharmatics
401 Terry Avenue N.
Seattle, WA 98109
Phone: (206) 802-6499
Fax: (206) 802-6411
E-mail: jason_johnson@merck.com

Samuel Karlin
Professor
Stanford University
Department of Mathematics
Stanford, CA 94305
Phone: (650) 723-2204
Fax: (650) 725-2040
E-mail: karlin@stanford.math.edu

Katherina Kechris
Post-doctoral Researcher
University of California, San Francisco
Department of Chemistry and Biophysics
600 16th Street, Room S441
San Francisco, CA 94143-2240
Phone: (415) 514-2617
Fax: (415) 514-4140
E-mail: kechris@genome.ucsf.edu

Chris Lee
Professor
University of California, Los Angeles
Molecular Biology Institute
Los Angeles, CA 90095
Phone: (310) 825-7374
Fax: (310) 267-0248
E-mail: leec@mbi.ucla.edu

Michael Levitt
Professor of Structural Biology and Chair
Department of Structural Biology
Stanford University School of Medicine
Room D109
Stanford, CA 94305
Phone: (650) 723-6800
Fax: (650) 723-8464
E-mail: michael.levitt@stanford.edu

Peter Li
Celera Genomics
45 W. Gude Drive
Rockville, MD 20850
E-mail: lipw@celera.com

Huiying Li
University of California, Los Angeles
Molecular Biology Institute
Box 951570
Los Angeles, CA 90095-1570
Phone: (310) 825-1402
Fax: (310) 206-3914
E-mail: huiying@mbi.ucla.edu

Mike Liang
Stanford University
300 Pasteur Drive MC 5120
Stanford, CA 94305-5120
Phone: (650) 725-8010
Fax: (650) 725-3863
E-mail: mliang@stanford.edu

Herb Lin
Senior Scientist
National Research Council
500 Fifth Street NW
Washington, DC 20001
Phone: (202) 334-3191
Fax: (202) 334-2318
E-mail: hlin@nas.edu

Shuo Liu
Stanford University
MSOB Room X215
251 Campus Drive
Stanford, CA 94305-5479
Phone: (650) 996-7820
E-mail: sliu@smi.stanford.edu

Stefano Lonardi
Assistant Professor
University of California
Department of Computer Science and Engineering
Riverside, CA 92507
Phone: (951) 827-2203
Fax: (951) 827-4693
E-mail: stelo@cs.ucr.edu

Itay Lotan
University of California, Berkeley
Donner Lab 472
Berkeley, CA 94720
Phone: (510) 486-6284
E-mail: itayl@cs.stanford.edu

Bruce A. Luxon
Professor
University of Texas Medical Branch
301 University Boulevard
Route 1157
Galveston, TX 77555-1157
Phone: (409) 747-6876
Fax: (409) 747-6850
E-mail: bruce@nmr.utmb.edu

Xiaotu Ma
University of Southern California
1429 W. 23rd Street
Los Angeles, CA 90007
Phone: (213) 740-2414
Fax:
E-mail: xma@usc.edu

Hanah Margalit
Associate Professor in Computational Molecular Biology
The Hebrew University of Jerusalem
Department of Molecular Genetics and Biotechnology
Hadassah Medical School
P.O.B. 12272
Ein Kerem Jerusalem, 91120
Israel
Phone: 972-2-6758614 / 6758647
Fax: 972-2-6784010 / 6757308
E-mail: hanah@md.huji.ac.il

Shipra Mehta
University of Southern California
1042 West 36th Place, DBR287
Los Angeles, CA 90095
Phone: (213) 740-2410
E-mail: shiprame@usc.edu

George L Gabor Miklos
Director
Secure Genetics Pty Limited
81 Bynya Road
Palm Beach
Sydney,
Australia
Phone: 61 2 9974-3000
Fax: 61 2 9974-3111
E-mail: gmiklos@securegenetics.com

Jan Mrazek
Stanford University
Department of Mathematics
Stanford, CA 94305-2125
Phone: (650) 723-2923
Fax: (650) 725-2040
E-mail: mrazek@stanford.edu

Mustumi Nakamura
Arizona State University
Department of Computer Science and Engineering
699 S. Mill Avenue
Tempe, AZ 85281
Phone: (480) 727-6047
Fax: (480) 965-2751
E-mail: mutsumi@asu.edu

Brian Naughton
Apartment 208, Building 94
1094 Tanland Drive
Palo Alto, CA 94303
Phone: (650) 213-8266
E-mail: briannau@stanford.edu

Jim Noyes
Professor
Wittenberg University
Box 720
Springfield, OH 45501
Phone: (937) 327-6142
Fax: (937) 327-7511
E-mail: jnoyes@wittenberg.edu

Debnath Pal
University of California, Los Angeles
Molecular Biology Institute #105
Los Angeles, CA 90095
Phone: (310) 206-3907
Fax: (310) 206-3914
E-mail: dpal@mbi.ucla.edu

Ranjan Perera
Associate Director
ISIS Pharmaceuticals
2292 Farraday Avenue
Carlsbad, CA 92008
Phone: (760) 603-2638
E-mail: rperera@isisph.com

Pavel Pevzner
Ronald R. Taylor Professor of Computer Science
University of California, San Diego
Department of Computer Science & Engineering
APM 3132
La Jolla, CA 92093-0114
Phone: (858) 822-4365
Fax: (858) 534-7029
E-mail: ppevzner@cs.ucsd.edu

Sylvia Plevritis
Stanford University
Radiology Department
Lucas MRSI Center, Room P267
1201 Welch Road, Mailcode 5488
Stanford, CA 94305
Phone: (650) 498-5261
Fax: (650) 723-5795
E-mail: sylvia.plevritis@stanford.edu

Vicente M. Reyes
Postdoctoral Fellow
University of California, San Diego
San Diego Supercomputer Center
9500 Gilman Drive, MC 0505
La Jolla, CA 92093-0505
Phone: (858) 822-3638
Fax: (858) 822-3610
E-mail: vreyes@sdsc.edu

Meenakshi Roy
Postdoctroal Fellow
University of California, Los Angeles
Room 601 Boyer Hall
611 Charles E. Young Drive
Los Angeles, CA 90095
Phone: (310) 794-4026
Fax: (310) 267-0248
E-mail: meenakshi@mbi.ucla.edu

Quansong Ruan
University of Southern California
Computational Biology
1042 W. 36th Place, DBR 289
Los Angeles, CA 90089
Phone: (213) 740-2409
Fax: (213) 740-2437
E-mail: ruan@usc.edu

Lukasz Salwinski
University of California, Los Angeles
205 Boyer Hall
Los Angeles, CA 90095
Phone: (310) 825-1402
Fax: (310) 206-3914
E-mail: lukasz@mbi.ucla.edu

Serge Saxonov
Stanford University
Beckman Center B403
279 Campus Drive
Palo Alto, CA 94305
Phone: (650) 723-5976
E-mail: saxonov@stanford.edu

Bradley K. Sherman
Director of Bioinformatics
Mendel Biotechnology, Inc.
21375 Cabot Boulevard
Hayward, CA 94545
Phone: (510) 259-6111
Fax: (510) 264-0254
E-mail: bsherman@mendelbio.com

Al Shpuntoff
IEEE CSB Conference
49 Showers Drive T408
Mountain View, CA 94040
Phone: (650) 208-8690
Fax: (650) 941-2015
E-mail: al@afs4dna.com

Elnat Spriznak
University of California, Los Angeles
105 Molecular Biology Inst.
Los Angeles, CA 90095
Phone: (310) 206-3907
Fax: (310) 206-3914
E-mail: elnat@mbi.ucla.edu

Nam Tran
Arizona State University
Department of Computer Science and Engineering
Tempe, AZ 85281-8809
Phone: (480) 921-1296
E-mail: namtran@asu.edu

Zhidong Tu
University of Southern California
Department of Computational Biology
1042 West 36th Place, DBR297
Los Angeles, CA 90089-1113
Phone: (213) 821-2231
Fax: (213) 740-2437
E-mail: ztu@usc.edu

Stella Veretnik
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537
Phone: (858) 534-8366
Fax: (858) 822-0873
E-mail: veretnik@sdsc.edu

Li Wang
University of Southern California
1042 W. 36th Place, DBR 202
Los Angeles, CA
Phone: (213) 821-3167
E-mail: wang7@usc.edu

Qi Wang
University of California, Los Angeles
609 Boyer Hall
Los Angeles, CA 90095
Phone: (310) 794-4026
Fax: (310) 267-0248
E-mail: wangqi@ucla.edu

Shoshana Wodak
Senior Scientist
The Hospital for Sick Children
Structural Biology and Biochemistry Program
555 University Avenue
Toronto, Ontario M5G 1X8
Canada
Phone: (416) 813-5724
Fax: (416) 813-5085
E-mail: shoshana@sickkids.ca

John C. Wooley
Professor
University of California, San Diego
700 University Center, Mail Code 0043
9500 Gilman Drive
La Jolla, CA 92093-0043
Phone: (858) 822-3604
Fax: (858) 822-4767
E-mail: jwooley@ucsd.edu

Yi Xing
University of California, Los Angeles
609 Boyer Hall
Los Angeles, CA 90095
Phone: (310) 794-4026
Fax: (310) 267-0248
E-mail: yxing@ucla.edu

Min Xu
University of Southern California
1042 W. 36th Place, DBR 202
Los Angeles, CA
Phone: (213) 821-3167
E-mail: mxu@usc.edu

Na Xu
University of California, Berkeley
367 Evans Hall
Berkeley, CA 94720-3860
Phone: (510) 791-6289
E-mail: naxu@stat.berkeley.edu

Hua Yang
University of Southern California
1042 West 36th Place, DBR287
Los Angeles, CA 90007
Phone: (213) 740-2410
Fax: (213) 740-2437
E-mail: huayang@usc.edu

Yuzhen Ye
The Burnham Institute
10901 North Torrey Pines Road
La Jolla, CA 92037
Phone: (858) 646-3100 x3634
Fax: (858) 713-9930
E-mail: yye@burnham.org

Iwei Yeh
Stanford University
Clark Center S242
318 Campus Drive
Stanford, CA 94305
Phone: (650) 725-8010
E-mail: iyeh1@stanford.edu

Xiaoyu Zhang
University of California, Los Angeles
Department of Molecular, Cellular and Developmental Biology
PO Box 951606
Los Angeles, CA 90095-1606
Phone: (310) 206-3336
E-mail: xiaoyu@plantbio.uga.edu

Kangyu Zhang
University of Southern California
1042 W. 36th Place, DBR 202
Los Angeles, CA 90007
Phone: (323) 363-5935
E-mail: kangyuzh@usc.edu

Keyan Zhao
University of Southern California
Molecular and Computational Biology Program
835 West 37th, SHS 172
Los Angeles, CA 90089
Phone: (213) 821-2819
Fax: (213) 740-8631
E-mail: kzhao@usc.edu

Lei Zhuge
Postdoctoral Fellow
University of Southern California
Computational Biology, DBR 289
1042 W. 36th Place
Los Angeles, CA 90089
Phone: (213) 821-2229
Fax: (213) 740-2437
E-mail: lzhuge@usc.edu