Overview
Research Summary: Protein Bioinformatics

As the 21st century unfolds, the biological sciences are being transformed by the advent of large-scale data. The sequencing of the human genome is a dramatic example of this. Simultaneous to this increase in biological data, computers and computation have had a transformative effect on the way information is handled, stored, and mined. These computational advances apply, of course, to many facets of life. The goal of my lab is to connect these two developments: harnessing computational advances for the analysis of large-scale biological data, principally by performing integrative surveys and systematic data mining.

More specifically, we are focused on protein bioinformatics: understanding the structure, function, and evolution of proteins through analyzing populations of them in databases and in whole-genome experiments. Overall we have four research foci, which follow a progression from surveying the overall genomic landscape to analyzing individual proteins and their interactions in more detail, to zooming in on the chemical structure of specific molecules.

1 Genomics: Mining and Annotating Intergenic Regions, especially in relation to Pseudogenes

We are involved in a number of large-scale collaborations (e.g. ENCODE) to probe the activity of intergenic regions with tiling array technology. We have developed tools to design, score and interpret these arrays and to highlight particular array artifacts. The overall conclusion from this work has been that much of the intergenic regions of the human genome appear to be active, both transcriptionally and in terms of protein binding. In connection with tiling array experiments, we have done an extensive amount of intergenic annotation, with a particular focus on mining intergenic regions for pseudogenes (protein fossils). We were, in fact, one of the first groups to perform comprehensive surveys of pseudogenes on a genome-wide scale in terms of protein families, which we did for human, worm, yeast and a number of other organisms. Collectively, our studies enable us to determine the common "pseudofolds" and "pseudofamilies" in various genomes and to address important evolutionary questions about the type of proteins that were present in the past history of an organism.
2 Proteomics: Using Networks to Mine Functional Genomic Data and Understand Protein Function

After the main elements of the human genome are identified, we need to characterize their function. We are trying to characterize gene function through molecular networks. We work on systematically integrating many weak functional genomic features with data mining techniques to predict protein networks (comprising protein interactions and other functional linkages). Some of the features integrated are obviously related to protein interactions (e.g. expression correlations), but many others such as gene essentiality are much less so. In addition, we have studied the structure of protein networks, both on a large scale in terms of global statistics (e.g. the diameter) and on a small scale in terms of local network motifs (e.g. hubs). In particular, we have correlated network hubs with gene essentiality. Most importantly, we extensively study the dynamics of networks. This has allowed us to show how a network dramatically changes in different conditions.

3 Structural Genomics: Analysis of Folds, Families and Functions on a Large Scale

Another area of research in our lab is structural genomics. Here, we conceptualize proteins not purely as character sequences or abstract network nodes, but more in terms of their molecular structure. We have examined the large-scale relationships between sequence, structure and function in order to understand the extent to which structural and functional annotation can reliably be transferred between similar sequences, particularly when similarity is expressed in modern probabilistic language. We have related the occurrence of protein folds and families to phylogeny and deep evolutionary history. Our studies enabled us to recognize that particular folds are more common in certain organisms than in others. Finally, as part of our work on structural genomics, we relate the properties of proteins with their eventual success at being purified and structurally characterized. This has been in the framework of a database and decision-tree mining framework that we have built for the NESG structural genomics consortium.
Computational Biophysics: Relating Macromolecular Motions and Packing

The final area of focus in the lab is analyzing small populations of structures in terms of their detailed 3D-geometry and physical properties. Here, we try to interpret macromolecular motions in terms of packing. We have set up a database of macromolecular motions and coupled it with simulation tools to interpolate between structural conformations; the database also has tools to predict likely motions based on simple models, such as normal modes and localized hinges connecting rigid domains. Part of this project involves devising a system for characterizing motions in a highly standardized fashion. Our motions classification scheme is motivated by the fact that protein interiors are packed exceedingly tightly, and the tight packing can greatly constrain a protein's mobility. We have developed tools for measuring and comparing the packing efficiency at different interfaces (e.g. inter-domain, protein surface, helix-helix, protein vs. RNA) using specialized geometric constructions (e.g. Voronoi polyhedra).

Summary & Broader Societal Issues

In summary, my lab acts a connector, bringing quantitative approaches from disciplines such as CS and applied math to bear on real questions and data in molecular biology. In particular, we have extensively applied classical computational approaches involving simulation, machine learning, and database design to biological problems. This often happens in the framework of practical, experimental collaborations, where we function as part of multi-disciplinary teams. Team participation is a key feature of the lab. Finally, as part of our mission to connect biology with computation, we have also extensively analyzed how a number of larger issues relating to computation in society impact biological research. In particular, we have examined how general aspects of e-publishing and digital libraries relate to biomedical databases and how various legal and security concerns significantly impact genomics database interoperation.
This is a research collaboration network centered on Dr. Mark Gerstein and Dr. Michael Snyder. Each eclipse stands for an individual researcher.
• Quantifying environmental adaptation of metabolic pathways in metagenomics.


• 2 figures
Strength of Pathway co-variation with environment

CCA structural correlation

Environmentally invariant

Environmentally variant

[ Gianoulis et al., PNAS (in press, 2009) ]
[ Gianoulis et al., PNAS (in press, 2009) ]
• PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.
• 1 figure
PeakSeq: Scoring Relative to Controls

- Filter for Potential Targets based on "Mappability" Simulation
- Scale Input Relative to ChIP
- Score Relative to Bionomial Expectation

[Scale Input: Pf = 0, Slope = 1.24, R² = 0.71]

[Enriched Sites]

[ChIP-Seq Sample]

[Potential Target Sites]

[Mappability Map]

[Potential Target Sites]

[Input DNA]

[Threshold]

[Score]

[Relative to]

[Bionomial Expectation]

[Filter for]

[Scoring]

[Relative to]

[Controls]
3

- Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution.
- 2 figures
Complexities in Pseudogene Annotation

Ribonucleoprotein A1 proc. pseudogene

ψHNRPA1

ψMTND2

ψMTND4

ψCYTB

Inserted mito. seq. resulting in 3 pseudogenes

(c) Mark Gerstein, 2002, Yale, bioinfo.mbb.yale.edu
### History of Pseudogene Preservation

Based on alignment from ENCODE MSA group


<table>
<thead>
<tr>
<th>Representative pseudogenes drawn from 201 total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>human</td>
</tr>
<tr>
<td>chimp</td>
</tr>
<tr>
<td>baboon</td>
</tr>
<tr>
<td>macaque</td>
</tr>
<tr>
<td>marmoset</td>
</tr>
<tr>
<td>galago</td>
</tr>
<tr>
<td>rat</td>
</tr>
<tr>
<td>mouse</td>
</tr>
<tr>
<td>rabbit</td>
</tr>
<tr>
<td>cow</td>
</tr>
<tr>
<td>dog</td>
</tr>
<tr>
<td>shrew</td>
</tr>
<tr>
<td>armadillo</td>
</tr>
<tr>
<td>elephant</td>
</tr>
<tr>
<td>tenrec</td>
</tr>
<tr>
<td>monodelphis</td>
</tr>
<tr>
<td>platypus</td>
</tr>
<tr>
<td>chicken</td>
</tr>
<tr>
<td>xenopus</td>
</tr>
<tr>
<td>tetraodon</td>
</tr>
<tr>
<td>zebrafish</td>
</tr>
</tbody>
</table>
A Bayesian networks approach for predicting protein-protein interactions from genomic data.


2 figures
• Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.
• 2 figures
All 44 ENCODE regions (29,998 kb)

Non-constrained

Constrained 4.9%

- Coding 32%
- UTRs 8%
- Other ENCODE experimental annotations 20%
- Unannotated 40%
6

• What is a gene, post-ENCODE? History and updated definition.
• 1 figure
Gene: An Evolving Concept

- The first appearance of the word gene derived from the Greek geneticoς or genesis.
- Alfred Sturtevant constructed the first genetic map.
- Griffith's experiment demonstrated the transformation of bacteria and suggested a principle of development.
- The classic example, demonstrated by Francis Crick and James Watson, is the DNA double helix structure.
- The DNA molecule was discovered by Watson and Crick in 1953.
- The Central Dogma of molecular biology was proposed by Francis Crick.
- The first human genome was sequenced in 1995.
- The first human genome was sequenced in 2007.
- The first large-scale gene function analysis using gene expression by panel.

Timeline:
- 1909: The first appearance of the word gene.
- 1910: The first genetic map constructed.
- 1953: The double helix structure of DNA was discovered.
- 1995: The first human genome was sequenced.
- 2007: The first large-scale gene function analysis using gene expression by panel.

In summary, the concept of the gene has evolved significantly over time, reflecting our understanding of genetic and biological processes.
7

- "Personal genomics requires redefining privacy The human blueprint: dangerous secrets"
- No figures
• Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library.
• 2 figures
Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match.
• Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context.
• 2 figures
Fig. 1. The human protein interaction network and its connection to positive selection. Proteins likely to be under positive selection are colored in shades of red (light red, low likelihood of positive selection; dark red, high likelihood) (6). Proteins estimated not to be under positive selection are in yellow, and proteins for which the likelihood of positive selection was not estimated are in white (6).
Fig. 2. Relationship of protein network centrality and single-nucleotide changes. (A) The periphery of the human interactome is strongly enriched for genes under positive selection. Shown is the correlation of the likelihood to be positively selected (6) and betweenness centrality (18). Dots are colored according to the same scheme as in Fig. 1. As expected for a highly significant Spearman rank correlation, almost all dots are near the x axis for high betweenness centralities, whereas high probabilities for positive selection are only observed at low betweenness centralities (Spearman $\rho = -0.06$, significant at $P = 1.2e-06$). (B) The periphery of the human interaction network is more variable on the protein sequence level. Shown is the ratio of nonsynonymous to synonymous SNPs vs. network centrality. A higher ratio (which corresponds to variability at the protein sequence level) tends to occur at the network periphery (Spearman $\rho = 0.1$, significant at $P = 4.0e-04$). (C Upper) Betweenness centrality of genes with some likelihood of being under positive selection (with a log-likelihood ratio >0) vs. all other genes. (C Lower) Betweenness centrality of genes with a high ratio of nonsynonymous to synonymous SNPs vs. genes with a low ratio of nonsynonymous to synonymous SNPs. The significance level of the differences is given as the Wilcoxon rank sum $P$ value between the bars.
10

- Relating three-dimensional structures to protein networks provides evolutionary insights.


- 5 figures
There remains a rich source of knowledge unmined by network theorists!

UTILIZING PROTEIN CRYSTAL STRUCTURES, WE CAN DISTINGUISH THE DIFFERENT BINDING INTERFACES
THAT IS HOW THE RESULTING NETWORK LOOKS LIKE

- **The Structural Interaction Network (SIN)**

  - Represents a “very high confidence” network
  - Total of 873 nodes and 1269 interactions, each of which is structurally characterized
  - 438 interactions are classified as mutually exclusive and 831 as simultaneously possible
  - While much smaller than DIP, it is of similar size as other high-confidence datasets

**Fig. 2.** Dependence of the average evolutionary rate ($dN/dS$ ratio) of a protein with the degree and the interacting accessible surface area (adjusted by protein size, as estimated from molecular weight). For the degree correlation coefficient, we get $r^2 = 0.05$, and for the adjusted interface surface area, $r^2 = 0.12$, suggesting that more than twice as much of the variation in $dN/dS$ is accounted for by adjusted interface surface area (12%) than by the degree (5%).
**Fig. 3.** The concept of network evolution by gene duplication. A given protein may acquire a new interaction by duplication of an existing one. Given equal likelihood of any gene to be duplicated, a protein with many partners is more likely to get a new partner than one with few—hence, there is effective preferential attachment. For single-interface hubs, this mechanism is straightforward. However, for multi-interface hubs, it would then require coevolution of the hub and the duplicated gene to form a new interface.
• Genomic analysis of the hierarchical structure of regulatory networks.
• 5 figures
Determination of "Level" in Regulatory Network Hierarchy with Breadth-first Search

I. Example network with all 4 motifs

II. Finding terminal nodes (Red)

III. Finding mid-level nodes (Green)

IV. Finding top-most nodes (Blue)

[Yu et al., PNAS (2006)]
Regulatory Networks have similar hierarchical structures

[S. cerevisiae]

[E. coli]

[ Yu et al., Proc Natl Acad Sci U S A (2006)]
Expression of MOT3 is activated by heme and oxygen. Mot3 in turn activates the expression of NOT5 and GCN4, mid-level hubs. GCN4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization.
Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers

B. Governmental hierarchy of a representative city (Macao)
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks

Average betweenness at each level

[Yu et al., PNAS (2006)]
• The Database of Macromolecular Motions: new features added at the decade mark.
• 5 figures
Example
"Morph": MBP

- 2 Known Crystal Structures (endpoints, not necessarily same seq.)
- Std. Geometric Stats. (from structure comparison)
- Pathway Interpolation

[Flores et al. (2006). NAR 34:D296.]
~19K morphs
(instances of conformational variability)
(384 canonical ones)

~200 classified motions

Motions collecting together and annotating Individual morphs into logical units

[Flores et al. (2006) NAR 34:D296.]
Adiabatic Mapping vs Linear Interpolation Strategies Compared with Calmodulinin
Transferrin hinge involves absence of steric constraints (continuously maintained interfaces), esp. at hinge