ATG V

"after the genome"

jackson hole, wyoming
october 6-10, 1999

ATG V

Experimental and Computational Methods to Understand Biological Function

What the workshop is about.

The postgenomic era is arriving faster than anyone had imagined. In fact, during 2000 we'll have a large fraction of the human genome sequence. Heretofore, our understanding of function has come from non-industrial experiments whose conclusions were largely framed in human language. The advent of large amounts of sequence data, and of "functional genomic" data types, such as mRNA expression data, have changed this picture. These data share the feature that individual observations and measurements are typically relatively low value adding. However, these data are now being generated so rapidly that the amount of information contained in it will surpass the amount of biological information collected by traditional means.

It is tantalizing to envision using genomic information to create a quantitative biology with a strong data component. Unfortunately, we are very early in our understanding of how to "compute on" this information so as to extract biological knowledge from it. In fact, current efforts to come to grips with genomic information often resemble a computer-savvy library science, where the most important issues concern categories, classification schemes, and information retrieval. When exploring new libraries, a measure of cataloging and inventory is surely inevitable. However, at some point we will need to move from library science to scholarship.

We would like to achieve a predictive understanding of biological function. To do this, we will need to recast our knowledge of biological systems in quantitative terms, in formulations in which that information becomes computable. We realize that making the bridge from knowledge about systems to the sets of abstractions that constitute computable entities is not easy. However, we believe that one aspect of these future formulations is now apparent; that they will need to be grounded in an understanding of their now-accessible molecular components. Completing the inventory of the identities and functions of these molecular components will require new experimental approaches. Creation of "computable" formulations that embody this knowledge may well make possible new computational approaches.

This year, After the Genome 5 will attempt to create a short-lived think tank. The goal is to suggest experimental and computational methods that will eventually lead to a quantitative and predictive understanding of biological function.

To this end, the workshop will bring together industrial and university researchers, including: 1) Physicists, chemists, and engineers who are devising and using new data gathering techniques, such as microarrays, protein mass spectrometry, and single molecule measurements, 2) Accomplished academic and industrial biologists, 3) Computer scientists from different subject areas who have experience moving from broad knowledge of systems to analysis of those systems at a level that enables that results in models and simulations, 4) Biologists and computer scientists who combine physiological experimentation and computer modeling to understand single cells and small networks of cells, 5) General visionary thinkers, and 6) Policy makers who can convey outcomes of the workshop to organizations that can commit resources.

Roger Brent

September 1999
Berkeley, California

ATG V

After the Genome V
Acknowledgements

Jackson Hole, WY
October 6-10, 1999

The vision and scientific program for this workshop are the work of Roger Brent, at the Molecular Sciences Institute (TMSI), Susan Burgess, of the Remuda Group, and Chris Sander, at Millennium Predictive Medicine. Susan Burgess also arranged overall conference logistics and corporate support.

The organizers would like to thank Finley Austin (Physiological Genomics) and Alain Rappaport (Carnegie Mellon University and NASA Ames) for great help in organizing policy and computational sessions and members of TMSI for useful discussions as to overall content.

The organizers would also like to thank Lauren Ha (TMSI) for coordinating support from the US governments, and Adrienne Regard (the Remuda Group) and Rebecca Hill (TMSI) for handling the many details of the meeting.

After the Genome V is sponsored by the Molecular Sciences Institute, in Berkeley, California. TMSI is a non-profit academic research institute. Its mission is to help create a predictive biology by integrating the development of computational methods with the development of new types of functional genomic and experimental information.

Roger Brent
Susan Burgess
Chris Sander

Jackson Hole
October 1999

ATG V

We are grateful for support from the following sources

US Defense Advanced Research projects Agency

US Department of Energy

US National Science Foundation

Dupont Pharmaceuticals, Inc.

Genetics Institute / Wyeth-Ayerst Research, Inc.

Millennium Predictive Medicine

The Molecular Sciences Institute

SmithKline Beecham, Inc.

Strategene, Inc.

ATG V

After the Genome V Schedule

Jackson Hole, WY
October 6-10, 1999

Wednesday, October 6, 1999

4:00 pm	Registration	Explorers Room
5:30 – 6:30 pm	Cocktails and Opening Dinner	

Introduction

Explorers Room

7:30 – 7:45 pm	Roger Brent, Molecular Sciences Institute *Introduction to the meeting: the problem of function*
7:45 – 8:30 pm	Bruce Damer, The Contact Consortium *After biological genomes*

8:45 – 10:00 pm	Reception	Trappers Room

Thursday, October 7, 1999

7:00 – 7:55 am	Breakfast	Homesteader Room

Morning Session

Explorers Room

Lessons from ways computers aided inquiry in other spheres

8:00 – 8:30 am	Andrew Moore, Carnegie Mellon University *New algorithms, architectures, and science for data mining of massive astrophysics sky surveys*

8:30 – 9:00 am Max Egenhofer, University of Maine
How spatial data is represented and computed in Geographic Information Systems: State of the art and beyond

9:00 – 9:30 am Break

Ways to computationally explore expression data

9:30 – 10:00 am John Weinstein, National Cancer Institute
Overview of clustering methods, clustered image maps (CIMs) and other data analysis tools for the Omic Era

10:00 – 10:30 am Greg Tucker-Kellog, Genetics Institute / Wyeth-Ayerst Research
Designed analysis of found experiments

10:30 – 11:00 am Orly Alter, Stanford
Singular value decomposition for processing and representing gene expression data

11:00 – 11:30 am Break

11:30 – 12:00 pm Michael Zhang, Cold Spring Harbor
Large-scale gene expression profiles and promoter analysis

12:00 – 12:30 pm Arturo Medrano, CIFN-UNAM
BClass: Bayesian self-organizing maps for clustering and classification of heterogenous biological databases

12:30 – 1:30 pm Lunch Homesteader Room

1:30 – 3:30 pm Recreation

Afternoon Session Explorers Room

Ways to computationally explore sequence data

3:30 – 4:00 pm Ross Overbeek, Argonne National Laboratory
Exploring genomes: strategic issues and the case for vehicular genomics

4:00 – 4:30 pm Christos Ouzounis, EBI-Hinxton
Protein interaction maps for complete genomes based on gene fusion events

4:30 – 5:15 pm Isidore Rigoutsos, IBM TJ Watson Research Center
and
Aris Floratos, IBM TJ Watson Research Center
The DELPHI pattern discovery program and its applications to biological, genomic, and text data

5:15 – 5:45 pm Stephen Rowley, Millennium Predictive Medicine
Latent Structure Indexing: letting the data tell you what the patterns are

6:30 – 7:30 pm Dinner Explorers Room

Evening Session Explorers Room

Computational structural biology

8:00 – 8:30 pm Tim Harris, Structural GenomiX
Structural genomic complements to functional genomic and proteomic approaches: assignment of biochemical function to novel proteins

8:30 – 9:00 pm Chris Sander, Millennium Predictive Medicine
Completing the map of the protein universe

9:00 – 9:30 pm Marvin Cassman, NIGMS
The structural genomics initiative

9:30 – midnight Reception Trappers Room

Friday, October 8, 1999

7:00 – 7:55 am Breakfast Homesteader Room

Morning Session Explorers Room

Technologies to approach function and structure

8:00 – 8:30 am John Welsh, UC San Diego

	Applying nRNA expression analysis to understanding cancer biology	
8:30 – 9:00 am	Katheryn Resing, U. of Colorado, Boulder *Analysis of signaling processes by functional proteomics*	
9:00 – 9:30 am	David Goodlett, University of Washington *Isotope distribution encoded tags for protein identification by constrained database searching with a single accurate peptide mass*	
9:30 – 10:00 am	Nancy Xu, Old Dominion University *Single-molecule analysis of ligand-receptor interactions*	
10:00 – 10:30 am	David Soll, University of Iowa *A computer-assisted system (3D-DIAS) for reconstructing, tracking and analyzing the substructures in a crawling cell*	
10:30 – 11:00 am	Break	

Simulation of biological systems

11:00 – 11:30 am	Dennis Bray, Cambridge University *Overview of methods for simulation of cell signaling pathways*	
11:30 – 12:00 pm	Dick Karp, UC Berkeley *Combinatorial problems in DNA microarray gene expression analysis*	
12:30 – 1:30 pm	Lunch	Deck of Explorers Room
1:30 – 2:30 pm	Recreation	

Afternoon Session Explorers Room

Getting the policies we need to understand biological function

Session Chair: Finley Austin

2:30 – 3:00 pm	Marvin Cassman and John Norvell, NIGMS *How NIGMS came to support structural genomics*

3:00 – 3:30 pm Laura Rodriguez, FASEB, Office of Public Affairs
How scientists can best interact with political leaders

3:30 – 4:00 pm Dick O’Neill, Highlands Forum
Methodologies for creating biology games and scenarios

4:00 – 4:20 pm Break

4:20 – 6:00 pm Panel Discussion: how do we get the policies we need

Panel Chair: Dick O’Neill

Panel Members: Richard O’Neill, Marvin Cassman, John Norvell, Laura Rodriguez, Finley Austin, Roxanne Ford

7:30 pm – Dinner and Honky Tonk, 1 Million Dollar Cowboy Bar and Restaurant, Jackson, Wyoming

Saturday, October 9, 1999

7:00 – 7:55 am Breakfast Homesteader Room

Morning Session Explorers Room

Interfaces between humans, information systems, and the natural world

8:00 – 8:30 am Alain Rappaport, Carnegie Mellon University
Artificial creatures exploring natural environments in space

8:30 – 9:00 am Rob Tow, Interval Research, Inc.
Computer literacy: the evolution of our interaction with the universe of information

9:00 – 9:30 am Brian Williams, MIT
Coordinating the regulatory and immune systems of robotic space probes and decentralized decision making robotic webs

9:30 – 9:50 am Break

9:50 – 10:20 am Nicola Muscettola, NASA Ames

The remote autonomous agent on Deepspace I

Can we use artificial life to explore biological questions?

10:20 – 10:50 am	Chris Winter, Cyberlife, Ltd. *Building virtual organisms: using computers to learn about biology*	
10:50 – 11:10 am	Break	
11:10 – 12:00 pm	Richard Lenski, University of Maine *Interactions between mutations in 'bugs' – digital as well as bacterial*	
12:00 – 12:30 pm	Bernhard Palsson, UCSD *Life on the edge: using genome-scale in silico models of microorganisms to interpret and predict metabolic phenotypes*	
12:30 – 1:30 pm	Lunch	Homesteader Room
1:30 – 3:00 pm	Recreation	

Afternoon Session Explorers Room

Advancing existing science and technology to further understand biological function

Workshop Session Chair: Susan Burgess

3:00 – 6:00 pm	Workshops leading to action items

Tentative Agenda Topics:

1. Draft an ATG-vision policy whitepaper
2. Prepare the post-workshop communication plan
3. Define an intentional structure-function strategy
4. Outline development paths for new-biology enabling technologies
5. Design specifications for an interactive pathway simulation
6. Define a difficult and significant biology problem for 2009 that uses the outputs of 1-5

7:30 pm –	Banquet	Mural Room

Sunday, October 10, 1999

Departure

8:00 – 10:30 am	Sunday brunch and discussions
10:30 – 2:00 pm	Recreation
2:00 – 5:30 pm	Depart

ATG V

After the Genome V
Abstracts

Jackson Hole, WY
October 6-10, 1999

Computational Genomics of Digital Organisms

Christoph Adami
California Institute of Technology
adami@krl.caltech.edu

Digital organisms are a new class of life form [1] which does not share a common ancestry with any other terrestrial form. Populations of self-replicating computer programs evolving and adapting *in silico* offer an opportunity to test generalizations about living systems that may extend beyond the organic life that biologists usually study. In particular, since digital orgamisms allow a complete characterization of the organism's function in terms of its sequence, it may constitute a useful test case for bioinformatic algorithms. In this talk, I present results of experiments on the interaction of mutations, and the role of complexity in maintaining robust genomes.

[1] R.E. Lenski, C.A. Ofria, T.C. Collier, and C. Adami, *Nature* 400 (1999) 661-667.

Singular Value Decomposition for Gene Expression Data Processing and Modeling

Orly Alter, David Botstein, and Patrick O. Brown
Department of Genetics and Biochemistry
Stanford University
orly@genome.stanford.edu

DNA microsrray technology and genome sequencing have advanced to the point that it is now possible to monitor gene expression levels on a genomic scale. These data hold the key to fundamental understanding of biological processes on the molecular level. The analysis of these new data requires mathematical tools which are independent of any gene expression model, suitable for making use of the large quantities of data and at the same time reducing the complexity of the data to make them comprehensible. Analysis so far is limited to clustering genes and arrays in order to identify groups of genes and arrays of similar expression patterns.

We describe the use of singular value decomposition in transforming gene expression data from genes/arrays space to “eigengenes”/”eigenarrays” space, where the eigengenes and eigenarrays are unique orthonormal superpositions of the genes and arrays, respectively. In this space the data are diagonalized such that each eigengene is expressed only in the corresponding eigenarray, with the corresponding eigenexpression level indicating their relative significance.

We show that several significant eigengenes capture most of the expression information, thus allowing for dimension reduction and interpolation of missing data. In some experiments, the significant eigengenes and eigenarrays can be identified as known independent processes, biological or experimental. Normalizing the data by filtering the prominent invariant eigengenes removes additive and multiplicative constants, which are known to be superimposed on the data by the experimental method, and enables meaningful comparison of the expression of different genes across different arrays in different experiments. Sorting the data according to the eigengenes and eigenarrays gives a global picture of the dynamics of gene expression, in which individual genes appear to be classified into groups of similar regulation and function, and individual arrays appear to be classified into groups of similar biological phenotype.

Computer Simulation of Cell Signaling Pathways

Dennis Bray
University of Cambridge
d.bray@zoo.cambridge.ac.uk

It is widely expected that computer analysis will be necessary before we can fully understand intracellular signaling pathways. But how best to perform this analysis? Conventional approaches using the numerical integration of continuous, deterministic rate equations can provide a convenient route to very large systems or those in which molecular details are not important. However, as the resolution of experimental techniques increases so the limitations of conventional models become more evident. Difficulties include the combinatorial explosion of large numbers of different species, the importance of spatial location and conformational changes in the communication process, and the instability associated with reactions between small numbers of molecular species.

A radically different approach is to represent each individual molecular species as a separate software object and then to apply Monte Carlo methods to predict the performance of the pathway. In such an approach, rate equations are replaced by individual reaction probabilities and the output has a physically-realistic stochastic nature. Techniques are available by which large numbers of related species can be codes in an economical fashion and key concepts, such as signaling complexes and the thermally-driven flipping of protein conformations, can be embodied into the program. Stochastic modeling may be the way forward that allows us to integrate biochemical and thermodynamic data relating to signal complexes into a coherent and manageable account.

NIGMS Support Programs for Structural Genomics

Marvin Cassman and John C. Norvell
National Institute of General Medical Sciences (NIGMS)
mc56j@nih.gov

norvell@nigms.nih.gov

In response to the growing interest in structural genomics, the NIGMS has sponsored three workshops during the past year to explore how best to support this emerging field. The first of these workshops examined the scientific goals of such an organized national effort in structural genomics as well as the feasibility of each of the constituent tasks. The second meeting focused on broad plans for this effort, with the recommendation that the Institute develop a support program for research centers. The third focused on protein target selection and cooperation between research groups and between international support agencies.

Following these meetings and further discussions with other scientists and with the Institute's Advisory Council, the NIGMS announced a new initiative in structural genomics. The structural genomics Request for Applications (RFA) calls for research centers that will each contain all the experimental and computational tasks of structural genomics. They will test strategies for high-throughput operations and serve as pilots for large-scale research networks of the future. The Institute plans to support three to six such research centers, each costing up to $3 million annually. The NIGMS also issued two Program Announcements (PAs) to encourage methodology and technology development in structural genomics. This support will come through individual research projects, program projects, and snall business innovation research projects. Background information, summaries of the workshops, and the RFA and PAs can be found on the NIGMS web site at: http://www.nih.gov/nigms/funding/psi.html.

Regional Scale Numerical Weather Forecasting and Visualization Going from Meteorological Data to Predictive Models

Zaphiris Christidis and Lloyd Treinish
IBM Thomas J. Watson Research Center
zaphiri@us.ibm.com

Numerical weather prediction models are in use today to predict weather operationally at relatively low resolution over a large geographic region (i.e. synoptic scale). Meteorologists employ such numerical models in combination with observed weather variables (temperature, wind speed and direction, humidity, etc.), to arrive at a final forecast. However, the resolution at which these models operate is often too coarse for the prediction of localized weather phenomena like thunderstorms, wind shear, land-sea breezes, etc. Weather forecasts can be substantially improved with the introduction of regional scale and mesoscale numerical modeling techniques. These regional models operate at higher resolution and incorporate explicit cloud microphysics. They do not replace the synoptic scale simulations, but supplement them by using the results to locally refine the weather simulations. In order to enable timely weather simulations, the Regional Atmospheric Modeling System (RAMS) has been parallelized on an IBM RS/6000 Scaleable Parallel (SP) distributed memory supercomputer. Since large volumes of complex data are being produced, the use of traditional graphical representations of data for forecasters can be burdensome. Instead, novel three-dimensional visualization strategies are employed. These methods are developed to provide a context for three-dimensional analysis, viewing and interaction. The complete system has been dubbed "Deep Thunder" and it has been used to enable reliable, affordable, high-resolution numerical weather prediction for a variety of applications. More information about Deep Thunder is available at http://www.research.ibm.com/weather. While operational weather forecasting has been the primary focus for the use of such weather modes, other potential applications can be considered beneficiaries of mesoscale weather simulations. These applications include travel, aviation, agriculture, broadcasting, energy, insurance and other industries where weather is an important factor for making effective business decisions.

Spatial Data Models in Geographic Information Systems: State of the Art and Beyond

Max J. Egenhofer
National Center for Geographic Information and Analysis
Department of Spatial Information Science and Engineering
Department of Computer Science
University of Maine
max@spatial.maine.edu
http://www.spatial.maine.edu/~max/

Spatial data models form the organizational foundation for geographic information systems. They typically separate geometric form semantic descriptions of data, and provide little support for temporal properties. A large variety of such spatial data models exist, allowing GIS users to capture spatial and query for geographic information in a variety of ways. Under the leadership of the Open GIS consortium and the ISO TC211 standards group, efforts are underway to streamline these diverse representations. This talk will review how geographic information is currently organized in GISs and discuss some of the mathematical models for expressing spatial relationships. It will also address modeling issues for the next generation of geographic information systems, such as the representation of qualitative spatial information, more powerful and comprehensive methods to deal with semantics, and new ways of making spatial similarity comparisons.

DELPHI: A Pattern-Based Method for Detecting Sequence Similarity

Aris Floratos, IBM
T.J. Watson Research Center
aris@us.ibm.com

We describe a new approach for identifying sequence similarity between a query sequence and a database of proteins. The central idea is the use of a set of patterns obtained from the underlying database through a one-time computation. These patterns are subsequently searched for on every query sequnce presented to the system. A pattern matched by a region of the query pinpoints to a potential local similarity between that region and all the database sequences also matching that pattern, In a final step, all such local similarities are examined more closely by alignin and scoring the corresponding query and database regions. By using a set of prudently chosen patterns, the tools presented in this work is able to discover weak but biologically important similarities. We provide a number of examples using both classified and unclassified proteins that corroborate this claim. Furthermore, the running time is much faster than any existing method since it depends on the size of the pattern set used and not on the underlying database. This last feature is of increased importance given the rate of accumulation of genomic data.

Isotope Distribution Encoded Tags for Protein Identification by Constrained Databse Searching with a Single Accurate Peptide Mass

David R. Goodlett*, James E. Bruce, Gordon A. Anderson, Beate Rist, Richard D. Smith and Ruedi Abersold
Department of Molecular Biotechnology
University of Washington
goodlett@u.washington.edu

Traditionally, protein sequences were determined by stepwise, chemocal degradation of purified or fragments thereof. With the advent of sequence database s which contain complete genomic sequences or large numbers of complete or partial expressed gene sequences (expressed sequence tags, EST's), the sequences of most proteins can be determined by correlating experimental data extracted from the protein with sequence databases. The many implemented sequence databse-searching strategies have in common the use of a combination of specific constraints to narrow down a candidate list of matching proteins in a databse to a single protein. Currently, the most restrictive constraints are generated by mass spectrometric (MS) or tandem mass spectrometric (MS/MS) analysis of peptide mixtures after proteolysis of a purified protein or protein mixture with a specific protease.

We describe a method for rapid and unambiguous identification of proteins by sequence database searching using the accurate mass of a single peptide and specific sequence constraints. Peptide masses were measured using Fourier Transform Ion Cyclotron Resonance-Mass Spectrometry (FTICR-MS) to an accuracy of 1 ppm. The presence of a cysteine residue within a peptide sequence was used as a database searching constraint. Cysteine-containing peptide were detected within a mixture of peptides by incorporating chlorine into a general alkylating reagent specific for cysteine residues. The natural isotopic distribution of chlorine encoded the cysteine-containing peptide with a distinctive isotopic pattern. The ability to identify proteins by the accurate mass of a single peptide represents a significant advance that should significantly speed proteome studies. Furthermore, the selective incorporation of an isotope-distribution encoded tag (IDEnT) into biomolecules is expected to find wide application in the analysis of mixtures where detection and isolation of analytes with specific structural features can be accomplished using high resolution mass spectrometry.

Structural Genomics Complements to Other Functional Genomic and Proteomic Approaches for the Large-Scale Assignment of Biochemical Function to Novel Proteins

Tim Harris
Structural Genomix
tim@stromix.com

Structural Genomics has come of age. Once the result of a lengthy and painstaking process, high-resolution three-dimensional protein structures can now be obtained with impressive speed and in unprecedented numbers. Using sequence databases, researchers can make judicious choices of polypeptide domains and their orthologues from several organisms. Expressing selenomethionine-labelled proteins in *E. coli* and other surrogate systems enables high-throughput crystallography and X-ray diffraction at many times current rates. Powerful database technology is available for both data capture and display.

While functional studies, particularly those that use genetics, give clear phenotypes, they do not necessarily provide any clues about precise biochemical function. Structural genomics complements existing functional genomics and proteomics approaches by implying biochemical function. For example, if protein A of unknown function has a similar structure to protein B for which there is a functional information available, then the two proteins may have the same or similar biochemical function.

For pharmaceutical and agricultural companies, early access to structural information about a given protein target offers great advantages in target selection. As many more real protein structures are solved, the power of molecular modeling to predict protein structure in silico will grow.

The San Diego-based company Structural GenomiX (a.k.a. SGX, formerly Protarch) was founded in 1998. SGX is creating a high-throughput structure determination platform to provide customers with structures on a commercial basis and to develop a unique database of the structures of hundreds of industrially relevant target proteins. SGX's proprietary structures will enhance the public domain initiatives in structural genomics by enabling scientists to draw additional implications about biochemical function from the currently limited data set.

The SNP Consortium Overview

Arthur Holden
The SNP Consortium
aholden@earthlink.net

The SNP Consortium has been formed to advance the field of medicine and the development of genetic based disgnostics and therapeutics, through the creation of a high quality, high density single nucleotide polymorphism (SMP map of the human genome, which will be made available to all parties at no cost. This presentation will summarize the history of this public initiative, its objectives, organizational structure, and the key technical approaches being utilized. The data release policies will be summarized to facilitate audience understanding of how this SNP database relates to other public data sources and can be most effectively utilized by all interested parties.

Combinatorial Problems in Gene Expression Analysis Using DNA Microarrays

Richard Karp
University of California, Berkeley
karp@cs.berkeley.edu

With the advent of DNA microarrays, data about the transcription levels of genes can be acquired far more efficiently than ever before. A single array experiment can measure the levels thousands of mRNAs. By measuring these levels under different experimental conditions one can observe the effects of different external conditions or gene knockouts and inductions on the functioning of cells. By measuring transcription in different tissue samples one can discover disgnostic tests for distinguishing normal tissue from neoplastic tissue.

The results of m array experiments on a set of n genes can be represented by a m x n matric of numbers. The i-j entry of the matrix gives the transciption level of the jth gene in the ith experiment. The experiments may be performed on different tissue samples, or on the same tissue samples or cell colony under different conditions, affected by temperature, time, growth conditions, drug treatments, gene knockouts and inductions, etc. A fundamental tool for mining this data is to perform clustering to partition the genes into sets of coregulated genes or to partition the experiments into sets of conditions with similar patterns of gene transcription. One can also go beyond traditional clustering to look for more refined patterns in the data; for example, certain sets of genes may behave similarly under certain experimental conditions, even though they are not coregulated under all conditions. We will describe some approaches to discovering such patterns of conditional coregulation.

One would like to use DNA microarrays to discover the structure of the pathways that regulate gene expression in cells. A pathway can be regarded as a dynamical system whose state includes the experimental conditions described above. A variety of mathematical models have been proposed for such pathways: the state variables can be treated as either discrete or continuous, the dynamics can be deterministic, nondeterministic or stochastic, and one can ne interested either in transient behavior or in steady-state behavior. We shall describe some initial work on the design of efficient experiments for inferring or verifying the structure of such pathways, and will discuss how this work might be extended to more realistic models.

Interactions Between Mutations in ‘Bugs’ – Digital as well as Bacterial

Richard E. Lenski
Center for Microbial Ecology
Michigan State University
lenski@pilot.msu.edu

Genetics research has usually proceeded by studying mutations one at a time, and considerable progress has been made that way. Of course, interactions among mutations have been studied, especially in the context of elucidating operons and other regulatory networks. Even so, we have little insight into the likelihood that arbitrary pairs of mutations will combine multiplicatively, or whether more complex interactions will occur. This lack of information becomes acute when we seek to compare whole genome sequences (in which pairs of even closely related organisms may differ by thousands of mutations) or to use DNA microarrays to infer genetic causality. We have performed experiments to examine the frequency of complex interactions among pais of random mutations in two different systems.

In *E. coli*, we used mini-Tn10 mutagenesis to produce random insertions and P1 transduction to construct genetypes that carried a pair of these mutations [1]. We measured the relative fitness of both single and double mutants, and we compared the observed fitness of each double mutant with the value predicted assuming no interactive effect. We saw an unexpectedly high frequency (~50%) of complex interactions, including both synergistic and antagonistic effects.

Digital organisms are computer programs that self-replicate, mutate, compete for CPU time, and adapt by natural selection [2]. We examined millions of mutations, alone and in combination, in 174 digital ‘species’, some of which have small genomes and can only self-replicate, whereas others have larger genomes and can perform complex logical functions as well as self-replicate. The more complex species were generally more robust with respect to both single and multiple mutations. In both simple and complex species, the majority of pairs of non-lethal mutations exhibit complex interactions with respect to overall performance of the digital organism.

[1] Elena, S.F. and Lenski, R.E. 1997. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* 390: 395-398.

[2] Lenski, R.E., Ofria, C., Collier, T.C., and Adami, C. 1999. Genome complexity, robustness and genetic interactions in digital organimsm. *Nature* 400: 661-664.

BClass: Bayesian self-organizing maps for clustering and classification of heterogenous biological databases

Arturo Medrano
CIFN-UNAM
amedrano@cifn.unam.mx

Motivated mainly by a problem of clustering and knowledge discovery in biological databases, classical mixture models are used to tackle the problem of multivariate, heterogeneous classification and clustering. Variables are described using various statistical distributions (categorical, discrete and continuous). A set of possible quite heterogeneous variables is tras=nsformed into a purely homogeneous set of “characteristics” represented by the probabilities for each entry to belong to each of the groups in the mixture, thus forming the stochastic matrix P. The (Bayesian) Inference then focuses on approximating P, that is, the grouping probabilities. This is done using Markov chain Monte Carlo methods (Metropolis-Hastings and Gibbs sampling algorithms).

The methodology is divided mainly in two stages:

1. Transformation of the original data set into P; a purely homogeneous set of variables with the same units (probabilities). To do so, it is important to define a number of groups (elements in the mixture) in order to represent each database entry as a stochastic vector with as many components as columns in the matrix P.

This stage is based on:

a) Statistical models for data representation.

b) Assumptions (e.g. conditional independence among data belonging to one group)

c) Bayesian statistical inference.

2. Results interpretation.

The output of BClass is the matrix P. This stochastic matrix must be interpreted in terms of the clusters that the entries form due to their intrinsic characteristics. So far, we have applied a number of techniques to visualize, define and interpret the clusters:

a) Graphical representation of the entries by means by means of “Archipelago” plots.

b) To visualize the probabilistic maximum for each entry, which tells us what group it is more likely that a given entry belongs to.

c) Once the clusters are defined, to obtain the variation ranges of each variable within each cluster.

The Computer Science of Data Mining, Anomly Hunting and Discovery in Massive Scientific Datasets

Andrew W. Moore
Carnegie Mellon University and
Schenly Park Research, Inc.
awm@cs.cmu.edu

This talk will describe the Auton Lab’s research project in Cached Sufficient Statistics, which has been applied to finding many kinds of patterns, clusters, functions, anomalies and statistical models in massive scale science databases from drug discovery, inventory management and the multi-terabyte Sloan Digital Sky Survey.

Conventional scaling up of traditional statistics, data analysis and pattern matching quickly runs into trouble as the datasets get large. For example, many statistic clustering algorithms would take weeks when applied to gigabyte-level databases.

Instead of resorting to supercomputers, I will show how software in the form of some “cached sufficient statistics” datastructures (some from the 1970’s, some newly developed) combined with new kinds of search algorithms can allow a user to do statistical analysis with millions to billions of records wit (almost) the speed and ease of current analysis on small datasets using spreadsheet or statistics packages.

If time permits I will show examples of this technology applied to clustering, density estimation, anomaly detection, biotoxin identification, medical diagnosis, very-high-dimensional correlation detection, Bayesian Network discovery, association rule learning, non-parametric statistics, and autonomous experiment.

Methodologies for Creating Biology Games and Scenarios

Dick O’Neill
Highlands Forum
rpon@bellatlantic.net

The purpose of this talk is to give members of various scientific disciplines a glimpse into how other communities of interest were able to bring disparate disciplines and agendas together to address common challenges in a productive manner. The examples of one cross-disciplinary body, the Highlands Forum, will be used as a case study, and I will describe the tools which we use to abstract key elements and use those abstractions to make an exercise that educated participants and illuminated strengths and weaknesses in the process.

Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events [Enright, et. Al., 1999, Nature, in press]

Christos Outzonis
EMBL
ouzounis@ebi.ac.uk

A large-scale effort to measure, detect and analyze protein-protein interactions with experimental methods is under way. These methods include biochemical, molecular biological and genetics approaches. Using the two-hybrid system, an international effort to analyze the complete yeast genome is in progress. All the above approaches are tedious, labor-intensive and inaccurate. From a computational perspective, the question is how can we predict that two proteins interact on the basis of structure or sequence alone. We present a novel algorithm that identifies gene fusion events in complete genomes, solely based on sequence comparison. By implication of the fact that there must be selective pressure for certain genes to be fused over the course of evolution, we are able to predict protein-protein interactions. We show that 215 genes/proteins in the complete genomes of *Escherichia coli*, *Haemophilus influenzae* and *Methanococcus jannaschii* are involved in 64 unique fusion events. The approach is general and applicable for genes of unknown function.

Exploring Genomes: the Strategic Issues

Ross Overbeek
Argonne National Laboratories
overbeek@mcs.anl.gov

The amount of available DNA sequence data has been doubling about every 18 months for over a decade. It appears that this exponential increase is sustainable for at least the next 5-10 years. This suggests that we will have on the order of 500-1000 genomes available for analysis within 5-6 years. There are deep disagreements about how one exploits such a collection; these disagreements relate to underlying goals which are often left unarticulated. An attempt will be made to present the alternative positions in a simple context, reasoning by analogy. I will then argue that:

1. Comparative analysis is the central theme, and that simultaneous analysis of hundreds of genomes is fundamentally easier than the analysis of single genomes.

2. There has been a significant advance in characterization of prokaryotic genomes (based on statistical analysis of clustering of genes on the chromosome). This, together with a number of other unrelated advances, will rapidly increase out understanding of the prokaryotic cell.

3. Characterization of eukaryotic genes and functional subsystems, even though driven by significantly different goals, will be limited by the same factors. Until large numbers of genomes become available, it will be difficult to formulate even an approximate understanding of the eukaryotic gene pool. With a large and diverse set of genomes, key problems become substantially easier to address.

Life on the Edge: Using Genome-Scale *In Silico* Models of Microorganisms to Interpret and Predict Metabolic Phenotypes

Bernhard O. Palsson
Department of Bioengineering
University of California, San Diego
bpalsson@ucsd.edu

Small genome sequencing and annotation are leading to the definition of metabolic genotypes in an increasing number of organisms. We show how *in silico* metabolic genotypes are formulated based on genomic biochemical and strain-specific data. Such metabolic genotypes have been formulated for *E. coli*, *H. influenzae*, and *H. pylori*. The *in silico* models are based on the philosophy of using applicable physico-chemical (such as stoichiometric structure) and capacity (maximum fluxes) constraints on the integrated functioning of the metabolic networks. Given these constraints, optimal phenotypes can be computed and compared to experimental data. They are found on the edge of the allowable solution spaces – where the governing constraint on cellular functions can be identified. For *E. coli*, this process leads to quantitative prediction of growth and metabolic by-product secretion data in batch, fed-batch, and continuous cultures, and to the accurate prediction of the metabolic capabilities of 73 of 80 mutants examined. Furthermore, we present mathematical methods that allow for the analysis, interpretation, prediction, and engineering of the metabolic genotype-phenotype relationship, and for the interpretation of expression array data.

Artificial Creatures Exploring Natural Environments in Space

Alain Rappaport
Carnegie Mellon University
atr@cs.cmu.edu

Deep space exploration requires the design of very innovative computing tools, aiming at a form of physical and reasoning autonomy while still interacting at various levels with other artifacts (including humans). Autonomy requires qualitative and quantitative models of spacecraft operations, with logic-intensive processes to carry out reasoning tasks in real-time. It also requires more cognitive, executive-level controls to manage tasks and priorities. On-board science is another key requirement, providing in situ data interpretation and enabling key decision-making capabilities. We have been exploring analogies between these and other computational approaches for autonomous or supporting creatures in space exploration and the acquisition and analysis of molecular and genomic information.

Analysis of Signaling Processes by Functional Proteomics

Katheryn A. Resing and Natalie G. Ahn
Department of Chemstry and Biochemistry and Howard Hughes Medical Institute
University of Colorado, Boulder
katheryn.resing@colorado.edu

Biochemical, molecular biological, and genetic studies have led to an increasingly complex picture of cell signaling mechanisms. However, cell phenotypes are typically assessed by assays of growth rate, cell shape, motility, adhesion, sensitivity to apoptosis, etc. Although changes in these properties are driven by changes in signaling pathways, it is impossible to directly or quantitatively connect these functional assays to specific signaling events. Furthermore, differences between cells can be subtle and difficult to measure. More precise ways of delineating the phenotype that are quantifiable and directly related to the underlying signaling status would be useful.

Our preliminary studies indicate that this can be accomplished by analyzing the cellular proteome (a proteome is the expressed proteins of a given cell type). Altered expression and/or post-translational modifications in the major proteins underlie the cell's phenotype, and these events are regulated by signaling-dependent transcriptional activation, message stability, and/or modifying enzymes. Thus, the proteome should reflect the signaling status of that cell type.

Using the immobiline dry-strip method for isoelectric focusing and two-dimensional gel electrophoresis (2D) and in-gel digestion of "spots" followed by identification of the proteins by mass spectrometry, we are carrying out four functional proteomics studies:

1. Analysis of specificity of a drug directed at the MAPK kinase pathway. These studies use a leulemia cell line that is induced to differentiate in response to phorbol ester or after transfection with constitutively active MAPK kinase and in cells activated. Exactly the same proteins are affected in the TPA induced cells by a drug directed towards MAPK kinase as are affected in response to the constitutively active kinase. This is an elegant way of testing a drug specificity, and also shows that the constitutively active transfection had no other effects on the cells.

2. Changes with tumor progression. Each metastatic cell has 38, 43, or 53 proteins that are significantly altered in expression, compared to a nonmetastatic melanocyte cell line. Furthermore, the three lines show a progressively altered proteome. We suggest that these are markers for the cumulative mutations in oncogenes, survival factors, or tumor suppressors that regulate tumor progression.

3. Transfection of cells with individual components of a "linear" pathway to demonstrate possible branch points. A comparison of cells transfected with constitutively active mutants of ras and MAPK kinase provides surprising new information on the extensive role of negative feedback in this pathway.

4. Analysis of differentiation. We are using an epidermal cell culture model to understand the diverse events that must be coordinated during extensive remodeling of this cell during terminal differentiation.

As the human genome project nears completion, our ability to move from the list of proteins that change to specific studies of post-translational modification or the promoter regions and mRNA stability determinants will improve. We anticipate that research will shift from emphasis on linear signaling pathways to paradigms that consider the entire signaling pattern and believe that this methodology will be an important driving force for that change.

Pattern Discovery and Its Applications to Biology and Elsewhere

Isidore Rigoutsos
Bioinformatics and Pattern Discovery Group
IBM T.J. Watson Research Center
rigoutso@us.ibm.com

One of the interesting problems in biological data analysis is the discovery of subsequences ("patterns") that are common to a given collection of related "streams". In the early days, the problem instantiated itself in the form of motif discovery in amino acid sequences. More recently, the concept has expanded into the discovery of patterns in structural data, gene expression data, scientific text, nucleotide sequences, gene marker expression data, etc. And new applications continue to be devised. In all such contexts, the discovered patterns reveal correlated elements that have an associated functional, structural or other significance.

In this talk, I will present and discuss algorithms and applications that we have developed in my group for carrying out pattern discovery in the bioinformatics context. These applications span a large spectrum: unsupervised motif discovery, multiple sequence alighnment, tandem-repeat discovery, gene expression analysis, functional annotation, local 3-dimensional structure characterization, and other. Also, specific examples taken from actual biological problems will be shown.

** This is joint work with the members of my group: Aris Floratos, Laxmi Parida, Yuan Gao, and Dan Platt.

Latent Semantic Indexing: Letting the Data Tell You What the Patterns Are

Stephen G. Rowley
Millennium Predictive Medicine
sgr@pobox.com

Indexing is the clustering of composite data objects based on their features, such as clustering books by title, author, or subject matter. Latent Semantic Indexing [e.g., Derrwester Dumais Furnas Landauer & Harshman 90] is such a technique based on singular-value decomposition of a feature-frequency matrix.

It is "semantic" because it uncovers meaningful features of the data, such as synonymy and polysemy: people interested in couches are likely to be interested in sofas as well, but the word "couch" can be used as an unrelated verb, as well. In this sense, it surpasses traditional lexical methods which make explicit mention of the tokens to be searched fore.

It is “latent” because the semantics are extracted automatically, by a process akin to principal components analysis. Rather than requiring a user to explain the categories of interest, the algorithm invents categories to interest based purely upon the data.

We give a lightly mathematical overview of the LSI process, illustrate a few examples, show its scales to large(ish) problems, and propose some applications to genome databases.

Completing the Map of the Protein Universe

Chris Sander
Whitehead Institute
MIT Center for Genome Research
Millennium Pharmaceuticals
sander@mpi.com

Within the next five to ten years, biologists will see an enormous increase in our knowledge about the sequence, function and structure of proteins. Several technical and organizational developments are responsible for this increase. Genome projects will deliver the sequences of all proteins from major species, including homo sapiens. New high-throughput experimental methods will yield rich information about the functional role of all proteins within the cell and for the organism. 3D structures from X-ray crystallography and nuclear magnetic resonance spectroscopy will be determined at a much faster rate.

It is therefore timely to undertake a comprehensive survey of all proteins: to map particular protein sequences and structures in the space of all proteins, to describe the evolutionary diversity of proteins of related function in different organisms and to develop predictive methods for rapid identification of function and structure from sequence information.

The presentation will focus on plans to determine a model of the 3D structure of all proteins, including tools for the selection of high value targets for structural genomics.

[1] Holm, L. and Sander, C. Mapping the protein Universe, Science 273 (1996) 595-602.

A Computer-Assisted System (3D-DIAS) for Reconstructing and Quantitatively Analyzing the Dynamic 3D Relationships of the Outer Surface, Nucleus, Pseudopods and Vesicles of the Crawling Cell

David R. Soll
W.M. Keck Dynamic Image Analysis Facility
The University of Iowa
dsoll@blue.weeg.uiowa.edu

As the genomes of less complex organisms are sequenced, and as the human genome emerges, out attention has focused on functional genomics, or how genes function to generate complex biological systems. The solution to function has, for the most part, been biochemical in nature. Pathways have been described and molecular interactions identified, but describing how these processes occur in space and time has been refractory, primarily because methods are not readily available for dynamic 3D reconstructions of molecular interactions on the grand scale of complex biological systems. One would think, however, that at a higher level of complexity such as the cell, which is readily visible through a variety of microscopy techniques, dynamic 3D reconstruction and analysis systems would be readily available and part of our everyday arsenal of cellular technologies. However, even at this macroscopic level of molecular interaction, methods for visualizing, reconstructing and analyzing dynamic events in 4D are scarce. In this presentation, one solution will be provided to the problem of macroscopic reconstruction and analysis.

Cell motility is fundamental to animal development, cellular immunity, maintenance of animal tissues in adult organisms, metastasis, disease progression and the survival of most single cell organisms. For that reason, understanding how cells locomote represents a pervasive theme in biology and medicine. Because most living cells are viewed through conventional microscopes from above or below, we tend to interpret both their cellular and intracellular behavior in two dimensions. However, static electron micrographs and confocal 3D reconstructions of fixed cells have demonstrated that they are indeed very three dimensional. Therefore, understanding how locomote must include 3D descriptions of behavior and dynamic architecture. To accomplish this, we have developed a computer-assisted 3D dynamic image analysis system, 3D-DIAS, that reconstructs the surface, nucleus and pseudopods of living crawling cells, and that tracks vesicle behavior in 3D. Using either differential interference contrast microscopy to visualize the cell surface, nuclear membrane, and pseudopodial regions, or a near-real time laser scanning confocal microscope to visualize vesicles in living cells, 30 optical sections are obtained in a one second period through the z-axis of a cell, and this procedure is repeated every second. The optical sections are digitized into the 3D-DIAS program, which image-processes each section, automatically outlines the edges of the nucleus and pseudopods, edges are manually entered. In the case of vesicles, 3D-DIAS automatically interprets position. 3D surfaces are encapsulated by faceting algorithms, and the final 3D image of the living, crawling cell as well as its nucleus, pseudopods and vesicles, can be viewed dynamically at any angle through a 3D stereo workstation. Over 100 parameters of motility and dynamic morphology are generated every second by 3D-DIAS for the cell surface, nucleus and pseudopods, and over 30 parameters of motility are generated every second for vesicles, providing the first high resolution, quantitative description of cellular and intracellular motility in 3D. This system has been used to identify behavioral defects in cytoskeletal and signal transduction mutatnts, HIV-infected T cells and HIV-induced T cell syncytia, normal and neoplastic cells, developing cardiac and embryonic cells, and embryonic organs. The audience will be supplied with 3D red and blue glasses, and provided with a computer-assisted demonstration of 3D reconstruction, and the 3D dynamics of the cell surface, nucleus, F-actin enriched pseudopods and vesicles of live, crawling *Dictyostelium* amoebae searching for and responding to chemotactic gradients of the attractant cAMP. Development of a quantitative 4D confocal system, a near real-time 3D reconstruction/motion analysis system, a high speed 3D system (performing complete 3D reconstruction every 0.1 sec) and a virtual reality system will also be briefly described.

Computer Literacy: the Evolution

Rob Tow
Interval Research, Inc.
robtow@tauzero.com

This talk will examine the evolution of “computer literacy”, meaning the ability to program computers, from its beginnings through its now near complete transformation in the general culture as tools for communication, and environments for social interaction, and almost no one programs them in the classical sense. Interval Research efforts in virtual reality environments and projective emotional communication with robotics will be used to illustrate some of the issues involved in this transformation.

Designed Analysis of Found Experiments

Greg Tucker-Kellogg
Genetics Institute
gtuckerkellogg@genetics.com

Experimental researchers in functional genomics are becoming rapidly overwhelmed by the volume of gene expression data generated from microarray experiments. Many of the questions relevant to these scientists are impossible to address from casual examination of the data, and researchers are still unsure how to handle what would be relatively simple questions in more traditional areas. I describe how statistical Design of Experiments (DOE) approaches can be used to drive efficient interpretation of expression data by treating each gene as a separate response variable. Examples will show how this approach complements other statistical questions of most interest.

Clustered Image Maps (CIMs) and Other Data Analysis Tools for the Omic Era

John N. Weinstein
Laboratory of Molecular Pharmacology
National Cancer Institute
weinstein@dptax2.ncifcrf.gov

Omic research (Weinstein, Science 282: 628, 1998) is fundamentally different from what molecular biologists did in the 45 or so years AD (After DNA) but BG (Before Genome). In the BG years, the young scientist typically encountered a gene, gene product, or process as a graduate student or fellow and then beat it to death for a career – using tools and concepts that hadn't been available when his or her mentor was beating it to death. Increasingly, however, biologists are doing omic research – surveying in aggregate the genes, proteins, or other molecular constituents of a cell, tissue, or organism. (Webster's dictionary defines "-ome" from New Latin as an "abstract entity, a group, or a mass", hence the term is etymologically and philologically respectable.) Omic research began, of course, with genomics. Then there were proteomics, kinomics, functional genomics, pharmacogenomics, CHOmics, immunomics, and metabolomics, *inter alia*. Clearly, pre-omics and omic modes of research should be considered as synergistic. The one provides necessary detailed information; the other provides context. Unfortunately, omic studies have been surprisingly difficult to fund and publish in an academic world obsessed with the narrow notion of "hypothesis-driven research". Biotech and pharma have caught on much quicker.

In some respects, howeverm the Omic Revolution is not bearing as much fruit as quickly as might be expected. One reason is that we face major challenges in the sheer size and, especially, the highly multivariate nature of the databases generated. Traditional forms of statistical analysis and data visualization don't tend to work very well in this domain. As a consequence, we and others have been developing a variety of tools based in part on classical statistics but also on modern computer-intensive statistical methods and artificial intelligence. In

my talk I will describe, among other methods, an extremely useful visualization technique that we introduced for omic research several years ago, the color-coded Clustered Image Map (CIM, or cc-CIM) [Weinstein, et. al., Stem Cells 12: 13, 1994 and Science 275: 343, 1997; Myers, et. al., Electrophoresis 18: 647, 1997]. Examples of data analysis and data visualization will be drawn from microarray-based gene expression studies that we and out collaborators at Stanford and the Whitehead Institute are doing on cancer cell lines in the National Cancer Institute's drug discovery program. Web-based programs for producing CIMs and doing other multivariate analyses can be found at http://discover.nci.nih.gov. Increasing numbers of such tools will be made available at the site in the coming months.

The RAP-Array Approach to cDNA Array Hrbridization

John Welsh
Sidney Kimmel Cancer Center
jwelsh@skcc.org

One of the persistent problems with current microarray technology is the difficulty of obtaining robust hybridization signals from rare transcripts. To address thus problem, optimization in dyes, detection systems, hybridization conditions, attachments strategies, and other variables are on going in many laboratories. An alternative solution involves methods for the construction of labeled cDNAs that reproducibly distort the representation of individual sequences. One such approach, amethod called RAP-Array, uses RNA arbitrarily primed-PCR fingerprinting or Differential Display to create probes with reproducibly altered abundances for individual sequences. Due to the selectivity of arbitrary priming, sequences in the low abundance-high complexity class are more highly represented in the resulting cDNA probe relative to transcripts in the higher abundance classes. Within a single sample, sequence abundances are highly distorted relative to the corresponding abundances in the original RNA population. However, this distortion is reproducible on a sequence-by-sequence basis, so that original differences in transcript abundance between two RNA sources are preserved.

In our RNA fingerprinting studies, we have encountered several interesting new phenomena. A new EGF and TGF-beta regulated gene, VAV3, and a natural 5'-trinkated variant, VAV3.1, will be discussed. Also, several phenomena have emerged from expression profiling of combinatoric treatment experiments, including a vector/vector complement pattern, a pattern characteristic of translation inhibition, and a particularly puzzling pattern that implies that UVC irradiation overrides the well-documented repressive effect that cycloheximide has on RNA turnover.

Coordinating the Regulatory and Immune Systems of Robotic Space Probes and Robotic Webs

Brian C. Wlliams
Massachusetts Institute of Technology
Space Systems and Artificial Intelligence Laboratories
williams@mit.edu

A new generation of sensor rich, massively distributed systems is emerging that offers the potential for profound economic and environmental impact, including building energy systems, deep space probes and sensor webs that monitor the earth ecosystem. These robotic webs have the richness that comes from interacting with physical environments, together with the complexity of networked software systems, They must be efficient, capable and long lived, that is, able to survive decades of autonomous operation within unforgiving environments. To achieve this level of performance software regulatory and immune systems must be developed that robustly coordinate senor and actuator activities internal to individual robotic systems, and that coordinate regulatory and immune systems of these robotic systems offers an overwhelming programming challenge. Traditionally programmers must reason through system wide actuators. The rsulting code typically lacks modularity, is fraught with error and makes severe simplifying assumptions.

Model-based autonomy meets this challenge through two ideas. First, we note that programmers generate the desired function based on their commonsense knowledge of how the software and hardware modules behave. The idea of model-based programming is to exploit this modularity by having engineers program reactive systems by simply articulating and plugging together these commonsense models. The second challenge is the unfeasibility of synthesizing a set of codes at compile time that envision all likely failure situations and responses. Our solution is to develop real time systems, called model-based executives that respond to novel situations on the order of hundreds of milliseconds, while performing extensive deduction, diagnosis and planning within their reactive control loop.

In this talk I will formulate a model-based executive as a deductive form of an optimal, model-based controller, in which are specified through a combination of concurrent, probabilistic transition systems and prepositional logic. This framework allows us to unify a diverse set of research results from model-based reasoning, planning, search, real-time propositional inference, and the theory of reactive languages. I will then discuss how reactivity is achieved using a high performance deductive kernal, called OPSAT that solves combinatorial optimization problems with constraints encoded in propositional logic. A first generation executive, called Linvingstone, was demonstrated this year on NASA’s first autonomous space probe, called Deep Space One, shortly before its asteroid encounter. Livingstone is also being demonstrated in a variety of space systems that include Mars rovers, Martian chemical plants, multi-spacecraft telescopes and the next generation shuttle. Finally, I will touch on future research that shifts from controlling the internals of single robotic systems to webs of robotic vehicles.

Building Virtual Organisms: Using Computers to Learn about Biology

Chris Winter
Cyberlife, Ltd.
chris.winter@cyberlifeco.uk

The last decade has seen the explosive growth of artificial life software as a setting of modeling tools. It has also seen a great rise in interest in computer scientists in trying to abstract from, understand and model biology. Sadly the lack of communication between the two communities has often meant such endeavors are seldom as powerful as would be desired.

The growing interest in highly realistic virtual worlds, containing thousands of agents that mimic plants, animals and people has led to the development of powerful modeling tools. The very complexity issues these problems can tackle are the same as those faced by biologists in trying to move the genome to the phenotype.

This talk will look at some of the developments in artificial life and examine whether the modeling techniques inspired there can help those working “After the Genome” and equally how biologists can help those building simulations and virtual worlds to get more powerful and realistic effects. The talk will concentrate on some of the philosophical and ways of thinking issues that prevent many people building large, scalable and complex models of biological organisms.

Single-Molecule Analysis of Ligand-Receptor Interactions

X. nancy Xu
Department of Chemistry & Biochemistry
Old Dominion University
chxu@odu.edu
http://www.odu.edu/~chem/xu.htm

Biochemical analyses at the single-molecule level present unique opportunities to study and characterize the chemical and physical properties of individual molecules. Potential applications in biochemical and biomedical research that are single-molecules analysis include tracking of individual steps in a sequence of biological events, early detection of biomarkers for diseases diagnosis and manipulating individual biological reactions. A variety of biological events inside the cells are induced by the interactions of ligands with recptors on the surface of cells (e.g., signal transduction, immune response). These biological cascades are associated with a variety of diseases (e.g., cancer, AIDS) from their onset and development through their diagnosis and treatment.

We are measuring real-time thermodynamic and kinetic parameters of single ligand-receptor interactions directly involved in immune regulation using real-time single-molecule fluorescence microscopy and spectroscopy. This study is an essential step in gaining an understanding of the initiation of diseases and in designing potential drugs. It also provides a unique opportunity to probe the initiation of cellular signaling pathways and fundamental theories (e.g., collision theory, lock-and-key model) at the single-molecule level. We are also studying affinity constants of ligand with receptor for the determination of co-receptors, antigen with antibody for the development of high selective immunoassays, and protein-protein interactions for the understanding of biological function and properties at the single-molecule level. For example, we have successfully conjugated T cell receptors (e.g., CD4, Fas) with ruthenium (II) tris-bipyridine that can be detected using both laser-induced fluorescence and electrochemiluminescence (ECL). We are able to monitor single receptors and single ligand-receptor complex using laser-induced fluorescence microscopy. We are tracking the interactions and dynamics of ligand with receptor and antigen with antibody in real-time at the single-molecule level. We have measured affinity constants and binding ratios of CD4 with gp120 and monoclonal antibody using ECL and compared the macroscopic measurements with single-molecule detection. We have also studied the competitive binding interactions of CD4 eith gp120, antibody and chemokines receptors (e.g., CCR5, CXCR4, CCR1) and studied the roles of chemokines (e.g., MIP-1 beta, MIP-1a, Human RANTES) in these binding interactions. This study will lead to the better understanding of co-receptors of HIV infection of T cell, chemokines therapy and HIV vaccine design. It may also lead to the development of ultrasensitive detection means for the earlier detection of disease and the discovery of new HIV co-receptors. The detailed experimental configuration and a variety of prospective applications will be discussed.

Large-Scale Gene Expression Profiles and Promoter Analysis

Michael Q. Zhang
Cold Spring Harbor Laboratory
mzhang@cshl.org

The use of high-density DNA arrays to monitor gene expression at a genome-wide scale constitutes a fundamental advance in biology. In particular, the expression pattern of all genes in Saccharomyces cerevisiae can be interrogated using microarray analysis where cDNAs are hybridized to an array of each of the ~6,000 genes in the yeast genome. In an effort to map all upstream regulatory elements, we started recently in collaboration with molecular biologists on developing theoretical and computational methods for the analysis of large-scale expression data. It is well known that complex gene expression patterns result from dynamic interacting networks of genes in the genetic regulatory circuitry. Hierarchical and modular organization of regulatory DNA sequence elements is important information for the understanding of combinatorial control and regulation of gene expression. I will use recent experimental data to illustrate how one can analyze expression profiles to extract regulatory cis-element information.

ATG V

After the Genome V Participants

Jackson Hole, WY
October 6-10, 1999

Stanley Abramowitz
Chemical & Biomedical Technology Office
NIST

Chris Adami
California Institute of Technology

Orly Alter
Department of Genetics & Biochemistry
Stanford University

M.J. Finley Austin
Physiological Genomics
Brigham & Women's Hospital

David Balaban
Affymetrix

Dennis Bray
Department of Zoology
University of Cambridge

Roger Brent
Molecular Sciences Institute

Jehoshua Bruck
California Institute of Technology

Ian Burbulis
Molecular Sciences Institute

Susan Burgess
The Remuda Group

William Busa
Cellomics, Inc

Lynn Caporale

Rob Carlson
Molecular Sciences Institute

Marvin Cassman
NIGMS

Zaphiris Christidis
IBM

Barbara Cohen
Nature Genetics

David Cohen
Burstein Laboratory

Alejandro Colman-Lerner
Molecular Sciences Institute

Bruce Damer
The Contact Consortium

Cynthia Edwards
Max Egenhofer
Department of Computer Science
University of Maine

Lee Eiden
NIMH

Drew Endy
Molecular Sciences Institute

Chris Fields
Perkin Elmer, Inc.

Aris Floratos
IBM T.J. Watson Research Center

Roxanne Ford
W.M. Keck Foundation

Rainer Fuchs
ARIAD

Guri Giaever
Stanford University

David Goodlett
Department of Molecular Biotechnology
University of Washington

Noelle Gracy
Academic Press

Michael Gruber

Lauren Ha
Molecular Sciences Institute

Tim Harris
Structural GenomiX, Inc.

Mariana Henkart
National Science Foundation

Rebecca Hill
Molecular Sciences Institute

Arthur Holden
The SNP Consortium

Richard Karp
Department of Computer Science
University of California, Berkeley

John Kellum
Asymetrix, Inc.

Eugene Kroll
Molecular Sciences Institute

Rajan Kumar
Sarnoff Labs

Bill Ladd
Spotfire, Inc.

Brenda Laurell
Interval Research, Inc.

Graham Lees
Academic Press

Han Lerach
Max Planck, Berlin

Richard Lenski
Center for Microbial Ecology
Michigan State University

Michael Liebman
Wyeth-Ayerst
Mary Lipton
Pacific Northwest National Laboratory

Larry Lok
Molecular Sciences Institute

Arturo Medrano
CIFN-UNAM

Eric Mjolsness
Jet Propulsion Laboratories

Andrew Moore
Carnegie Mellon University

Rick Norgren
Norgren Systems

John Norvell
NIGMS

Richard O'Neill
Highlands Forum

Christos Ouzounis
EBI Hinxton
EMBL

Ross Overbeek
Argonne National Laboratories

Bernhard Palsson
Department of Bioengineering
University of California, San Diego

Ljiljana Pasa Tolic
Pacific Northwest National Laboratory

Dhiraj Pathak
SmithKline Beecham
Bill Pullybank
IBM

Alain Rappaport
Department of Computer Science
Carnegie Mellon University

Adrienne Regard
The Remuda Group

Katheryn Resing
Department of Chemistry & Biochemistry
HHMI, University of Colorado, Boulder

Isidore Rigoutsos
Bioinformatics & Pattern Discovery Group
IBM T.J. Watson Research Center

Rose Ritts
Sarnoff Labs

Laura Rodriguez
Office of Public Affairs
FASEB

Stephen Rowley
Millennium Predictive Medicine

Chris Sander
MIT Center for Genome Research and
Millennium Predictive Medicine

Rimli Sengupta
University of Washington

Jim Schwaber
Dupont Computational Neuroscience
University of Pennsylvania

Victoria Smith
Genentech

Lydia Sohn
Princeton University

David Soll
W.M. Keck Dynamic Image Analysis
University of Iowa

Joseph Sorge
Strategene

Marvin Stodolsky
U.S. Department of Energy

Rob Tow
Interval Research, Inc.

Greg Tucker-Kellog
Genetics Institute

John Weinstein
Laboratory of Molecular Pharmacology
NCI

John Welsh
Sidney Kimmel Cancer Center

Brian Williams
Space Systems & Artificial Intelligence
Laboratories
MIT

Rusty Williams
Chiron

Chris Winter
Cyberlife, Ltd.

Jeff Wiseman
SmithKline Beecham

Barbara Wold
California Institute of Technology

Gordon Wong
Genetics Institute

Nancy Xu
Department of Chemistry & Biochemistry
Old Dominion University

Michael Q. Zhang
Cold Spring Harbor Laboratory