SWISS-PROT in the 21st Century!

Amos Bairoch

Swiss Institute of Bioinformatics, Geneva, Switzerland.
bairoch@sib-sib.ch www.expasy.ch/www/amos.html

SWISS-PROT is a curated added-value protein sequence database that strives to provide a high level of annotations (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy, and a high level of integration with other databases [1]. It currently contains about 80,000 annotated sequence entries from 6,500 species. It is used by an estimated 200,000 users worldwide and accessed through many different distribution media—the most popular one being currently the world wide web (see www.expasy.ch/sprot/). SWISS-PROT is complemented by TrEMBL, a computer-annotated supplement.

We will describe what we believe are the challenges that face SWISS-PROT in the near and not so near future. How can knowledge provided by scientists cohabit with computerized inferences? How will databases, journals, and researchers develop new relationships? What will be the impact of proteomic projects on the characterization of proteins? These are a few of the many questions that we will attempt to briefly evoke.

Reference


The Origin of Biological Information

Manfred Eigen

Max-Planck-Institut für biophysikalische Chemie,
Göttingen, Germany

What is the distinguishing feature of a living system that singularizes it from every non-living chemical ensemble, regardless of the extent of the complexity? The differentiable characteristic of the living system is information. Information assures the controlled reproduction of all the constituents, thereby ensuring the conservation of viability. Information—unlike energy—is not subject to a conservation law. Hence the fundamental question behind the origin of life is: How can information originate?

Information theory, which was pioneered by Claude Shannon, cannot answer this question. This theory is most successful in dealing with problems of coding and transmission. In principle, the answer was formulated 130 years ago by Charles Darwin: The information that is unique for life evolves by virtue of natural selection. Today we can be more specific: Natural selection is a non-equilibrium process. It is an inherent consequence of mutagenous self-replication at several levels of organization: for instance it is evident in molecules such as nucleic acids, in molecular complexes such as viruses, and in autonomous forms of life such as micro- or higher organisms. New physical concepts have been introduced in order to deal quantitatively with the dynamics of the molecular generation of genetic information. They provide a physical foundation for Darwinian behavior, yet they introduce major modifications in its interpretation. The lecture deals with these physical concepts, such as "sequence space," "quasi-species," and "hypercycles," and will scrutinize their adequacy for rationalizing experimental results obtained with molecular model systems and with viruses under natural conditions. Elucidating the principles of molecular self-organization has made possible the construction of automated machines that make it possible for genetic information to evolve under controlled conditions in an abridged time scale.

References

With the advent of DNA microarrays, data about the transcription rates of genes can be acquired far more efficiently than ever before. A single array experiment can measure the levels of 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 diagnostic 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* matrix of numbers. The *i,j* entry of the matrix gives the transcription level of the *j*-th gene in the *i*-th experiment. The experiments may be performed on different tissue samples, or on the same tissue sample 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. We will describe several different approaches to these clustering problems. One can also go beyond 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 abundancies of certain mRNAs and proteins, and whose inputs include 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 be 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.

This talk represents joint work with many colleagues at the University of Washington and other institutions in the Seattle area.

The field of genomics was radically changed with the sequencing of the first complete microbial genome, *Haemophilus influenzae* by The Institute for Genomic Research (TIGR) [1]. This project made it apparent that the DNA of entire complex organisms many megabases in size could be accurately and rapidly sequenced by using a “shotgun” sequencing strategy. Since that time, TIGR and other labs have combined to completely sequence the genomes of over 20 microbes. Knowing the complete genome sequence of the pathogens in this group will open up exciting opportunities to develop novel pharmaceuticals, biologics, and vaccines. The genomes of two important eukaryotic model organisms, *S. cerevisiae* [2] and *C. elegans* [3] have also been completed. In addition, several chromosomes from *P. falciparum* and *A. thaliana* are finished and these entire genomes will soon be complete.

Across all of these species, nearly half of the candidate genes that have been identified cannot be assigned a definitive biological role, leaving open a tremendous opportunity for functional as well as computational genomics. On the other hand, by a combination of molecular sequence analysis techniques, new insights have been made concerning the metabolic pathways, cell-surface receptor and transporter complement, and phylogeny of these organisms. The availability of these complete genomes makes comparative genomic analysis possible, leading to the discovery of synteny among organisms as well as regulatory and developmental networks controlling the expression of genes. The integration and semantic representation of this wealth of data will be critical to our ability to understand it.

At Celera Genomics we have set our goal to become the definitive source of genomic and associated medical information that will be used by scientists to develop a better understanding of the biological processes in humans and agriculturally important organisms and deliver improved health care in the future. Using breakthrough DNA sequencing technology, we are operating a genomics sequencing facility with an expected capacity greater than that of the current combined world output[4]. The early focus at Celera will be on completing the genomes of human, mouse, Drosophila, and rice. While the size of these genomes and the speed with which they will be sequenced will present enormous computational challenges for the discovery and characterization of genes, they represent an enormous opportunity to advance the complete understanding of living systems.
tire genes are complemented by numerous intragenic that horizontal transfer and lineage-specific loss of en-
zontal gene transfer and gene loss. These studies show
narios that include a number of distinct events of hori-
possible to construct parsimonious evolutionary sce-
DNA repair and programmed cell death, it is now
occasions, in detailed studies of protein super families
lows us to delineate families of orthologs across a wide
 Computer analysis of complete genomes of unicellu-
organisms shows that protein sequences are in gen-
eral highly conserved in evolution, with at least 70% of
epochs are currently available, and many more are in the pipeline. Considerable comparative analysis of these
already been performed, and while even more challenging work lies ahead, it is fair to ask at this
juncture, what is the impact of this research on biology in general. In my opinion, comparative analysis
complete genome has already affected our ideas of what biological evolution is to such an extent that it is
appropriate to claim a paradigm shift in evolutionary biology.
Computer analysis of complete genomes of unicellu-
ary organisms shows that protein sequences are in gen-
eral highly conserved in evolution, with at least 70% of
of them containing ancient conserved regions. This al-
many more are in the pipeline. Considerable comparative analysis of these genomes has already been performed, and while even more challenging work lies ahead, it is fair to ask at this
...
The Human Genome Project is providing life science researchers with access to unprecedented amounts of raw sequence data. To effectively harness this data and apply it to biomedical research, therapeutic development, clinical practice, and patient management, powerful new tools for measuring gene expression, polymorphism discovery, and genotyping are needed. GeneChip® probe arrays are powerful tools to meet these requirements. Light-directed chemical synthesis is used to generate miniaturized, high-density arrays of oligonucleotide probes called GeneChip probe arrays. Application specific oligonucleotide arrays have been used to rapidly scan known genes and discover genetic variants, to detect the presence of known alternative alleles, and to simultaneously measure the expression of thousands of individual genes. An integrated GeneChip® system including instrumentation and software has been developed for array hybridization, fluorescent detection, and data acquisition and analysis. Experiments demonstrating the effectiveness of these methods of genetic analysis will be described as well as new bioinformatics challenges generated by the new information.

References may be found at www.affymetrix.com/technology/papers.html

Gene Function via the Mass Spectrometric Analysis of Multi-Protein Complexes

Matthias Mann

Protein Interaction Laboratory, University of Southern Denmark—Odense University, Odense, Denmark. (www.pil.sdu.dk)

The anticipated availability of virtually all human gene sequences already within a year will usher in the “post-genome era” of biology sooner than expected. We now require large-scale experimental approaches which will use the genomic information but add another dimension of information to it. Methods which are already being applied include large scale two hybrid screening (currently for small to medium genome sizes) and large scale expression analysis via DNA chip arrays. Here we discuss an additional approach which is also capable of providing function or at least the cellular role of the genes uncovered in genomic sequencing projects. Advances in mass spectrometry over the last few years now make it possible to identify large numbers of gel separated proteins at minute levels (low femtomole/low nanogram) [4], [5]. Proteins of interest can be precipitated using gene tagging or antibody methods, revealing interacting proteins on one or two dimensional gels which can then be identified by mass spectrometry [1],[2]. We show that this technology can be scaled up to large numbers and that significant biological results have already been obtained both in structural protein complexes and in transient complexes such as the ones involved in signaling [3], [6]. In principle this technology can lead to a protein interaction map of the cell. The approach should be accompanied by bioinformatics tools which interpret the empirically found interactions. We conclude that mass spectrometry of multi-protein complexes is a valid approach which rapidly yields functional information on open reading frames identified in sequencing projects.

References


Exploiting Protein Structure in the Post-Genome Era

Michael J. E. Sternberg
Paul A. Bates, Lawrence A. Kelley, Robert M. MacCallum, Arne Müller, Stephen Muggleton, Marcel Turcotte


Diverse and innovative computational approaches are required to exploit the information encoded in protein structures so the knowledge can be used to interpret the explosion of genome sequence data. In particular, algorithms are required to predict protein structure and function from sequence. To illustrate the computational challenges, the following topics currently being considered in our laboratory will be described:

- The need to encapsulate expert knowledge in protein structure prediction
- The strategy to assign protein folds to genome sequences
- The detection of remote protein homologues using information from protein structures
- The use of inductive logic programming, a branch of machine learning, to identify principles of protein folding.