

Invited Talks

SWISS-PROT in the 21st Century!

Amos Bairoch

Swiss Institute of Bioinformatics, Geneva, Switzerland.
bairoch@isb-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

- [1] Bairoch A. and Apweiler R., *Nucleic Acids Res.* 27: 49-54(1999)

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 formu-

lated 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

- [1] Eigen, M.; McCaskill, J.; and Schuster, P. (1988), Molecular quasi-species, *J. Phys. Chem.*, 92, 6881-6891.
[2] Eigen, M.; McCaskill, J.; and Schuster P. (1989), The molecular Quasi-species, *Adv. Chem. Phys.*, 75, 149-263.
[3] Leuthusser, I. (1986), An exact correspondence between Eigen's evolution model and a two-dimensional Ising system, *Chem. Phys.*, 84, 1884-1885.
[4] Eigen, M. (1993), The origin of genetic information: viruses as models, *Gene*, 135, 37-47
[5] Eigen, M. and Nieselt-Struwe, K. (1991), How old is the immunodeficiency virus?, *AIDS*, 5, 585-593.
[6] Eigen, M. and Gardiner, W. C. (1984), Evolutionary molecular engineering based on RNA replication, *Pure Appl. Chem.*, 56, 967968.
[7] Eigen, M.; Biebricher, C. K.; Gebinoga, M.; and Gardiner, W. C. (1991), The Hypercycle—Coupling of RNA and Protein Biosynthesis in the Infection Cycle of an RNA-Bacteriophage, *Biochemistry*, 30, 11005-11018.
[8] Brakmann, S.; and Eigen, M. (1996), Evolution in the Test Tube, *Frontiers in Biology*, Vol. 1 (eds. W. Gilbert and G. Tocchini, Valentini) submitted.
[9] Eigen, M. and Rigler, R. (1994), Sorting single molecules: Application to diagnostics and evolutionary biotechnology, *Proc. Natl. Acad. Sci. USA*, 91, 5740-5747.
[10] Oehlenschläger, F.; Schwillie, P.; and Eigen, M. (1996), Detection of HIV-1 RNA by nucleic acid sequence-based amplification combined with fluorescence correlation spectroscopy, *Proc. Natl. Acad. Sci., USA*, 93.

Combinatorial Problems in Gene Expression Analysis Using DNA Microarrays

Richard M. Karp

University of Washington, Seattle, WA, USA

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 $m \times n$ 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.

Computational Genomics: Biological Discovery in Complete Genomes

Anthony R. Kerlavage

Celera Genomics, Rockville, MD, USA

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.

References

- [1] Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. Fleischmann, R. D. et al. *Science* 269: 496-512, 1995.
- [2] The Yeast Genome Directory. *Nature* 387(Suppl): 5-105, 1997.
- [3] Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. The *C. elegans* Sequencing Consortium. *Science* 282: 2012-2018, 1998.
- [4] Shotgun Sequencing of the Human Genome. Venter, J. C.; Adams, M. D.; Sutton, G. G.; Kerlavage, A. R.; Smith, H. O.; and Hunkapiller, M. *Science* 280: 1540-1542, 1998.

Comparative Genomics: Is It Changing the Paradigm of Evolutionary Biology?

Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

About 20 complete genome sequences of cellular life forms—bacteria, archaea, and eukaryotes—are currently available, and 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 juncture, what is the impact of this research on biology in general. In my opinion, comparative analysis of 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 unicellular organisms shows that protein sequences are in general highly conserved in evolution, with at least 70% of them containing ancient conserved regions. This allows us to delineate families of orthologs across a wide phylogenetic range and in many cases, predict protein functions with reasonable confidence. Once a robust set of such orthologous families is established, it is possible to examine the pattern of phylogenetic representation (or for brevity, simply a phylogenetic pattern) for each of them. Such an examination readily shows that for the great majority of orthologous families, the phylogenetic distribution is quite patchy, and in many cases, unexpected. Only ~100 families, most of which include components of the translation machinery, are universally conserved in all sequenced genomes. These observations indicate that horizontal gene transfer and lineage-specific gene loss are not inconsequential evolutionary quirks but rather prevailing forces of evolution, at least in the prokaryotic world. On many occasions, in detailed studies of protein super families and even entire functional systems, such as those for DNA repair and programmed cell death, it is now possible to construct parsimonious evolutionary scenarios that include a number of distinct events of horizontal gene transfer and gene loss. These studies show that horizontal transfer and lineage-specific loss of entire genes are complemented by numerous intragenic

recombination events that manifest in domain rearrangement at the protein level. Previously, such rearrangements were associated primarily with exon shuffling but the analysis of complete genomes shows that they are critically important also in the prokaryotic world where this mechanism is not operative.

Examination of phylogenetic patterns for families of orthologous proteins also results in more specific conclusions some of which may have far-reaching consequences. In particular, it is now clear that the basic DNA replication machineries (that is, the replicative DNA polymerases, primases, helicases, and several other proteins) in bacteria and in archaea/eukaryotes are *not* orthologous and may have evolved independently. This is in sharp contrast with the remarkable conservation of the components of the translation apparatus and the core transcription machinery. Other components of the DNA replication system, such as the sliding clamp (PCNA) and the clamp-loader ATPase, however, are universal. Apparently the most parsimonious, even if somewhat unconventional, interpretation of these observations is that the common ancestor of all extant cellular life forms (the so-called cenancestor) did not possess a modern-type, DNA-based replication and expression system although it did encode advanced translation and transcription machineries and a considerable repertoire of metabolic enzymes. Instead of a dsDNA genome, the cenancestor might have had a mixed system of small RNA and DNA genetic elements that were inter-converted via cycles of transcription and reverse transcription. This model seems to be able to account for both universal and distinct components of the DNA replication machinery in bacteria and archaea-eukaryotes.

Further genome sequencing, particularly of genomes of deep-branching organisms, will put these concepts to test and in any case, will add more substance for critical analysis. This must be complemented by further developments of methods for theoretical and ultimately experimental analysis of evolution on the basis of multiple genome comparison.

Acknowledgments

I am grateful to L. Aravind and Detlef Leipe for numerous stimulating discussions, without which these ideas could not have been developed in a clear form. I am aware and appreciative of the major contribution of Carl Woese to our understanding of the possible nature of the Universal Ancestor (Woese, 1998).

References

- [1] Aravind, L.; Tatusov, R. L.; Wolf, Y. I.; Walker, D. R.; Koonin, E. V. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* 14: 442-444, 1998.
- [2] Jeffares, D. C.; Poole, A. M.; Penny D. Relics from the RNA world. *J. Mol. Evol.* 46: 18-36 (1998).
- [3] Koonin, E. V.; Mushegian, A. R.; Galperin, M. Y.; Walker, D. R. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25: 619-637, 1997.
- [4] Makarova, K. S.; Aravind, L.; Galperin, M. Y.; Tatusov, R. L.;

- Wolf, Y. I.; Koonin, E. V. Comparative genomics of the archaea: universal and unique protein families. *Genome Res.*, in press, 1999.
- [5] Mushegian, A. R. and Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA.*, 93: 10268-10273, 1996.
- [6] Poole, A. M.; Jeffares, D. C., Penny D. The path from the RNA world. *J. Mol. Evol.* 46, 1-17 (1998).
- [7] Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. A genomic perspective on protein families. *Science*, 278: 631-637, 1997.
- [8] Woese, C. The universal ancestor. *Proc. Natl. Acad. Sci. USA* 95: 6854-6859 (1998).

Genes, Chips, and Genomes

Robert J. Lipshutz

Affymetrix, Inc., Santa Clara, CA, USA

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

- [1] Lamon, A. I. and Mann, M. (1997). Cell Biology and the Genome Projects—a concerted strategy for characterizing multi-protein complexes using mass spectrometry. *Trends in Cell Biology* 7: 139-142.
- [2] Neubauer, G.; Gottschalk, A.; Fabrizio, P.; Séraphin, B.; Lhrmann, R.; and Mann, M. (1997). "Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by Mass Spectrometry." *Proceedings of the National Academy of Sciences USA* 94: 385-390.
- [3] Neubauer, G.; King, A.; Rappsilber, J.; Calvio, C.; Watson, M.; Ajuh, P.; Sleeman, J.; Lamond, A. I.; and Mann, M. (1998). Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet* 20: 46-50.
- [4] Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie H.; and Mann, M. (1996). Linking Genome and Proteome by Mass Spectrometry: Large Scale Identification of Yeast Proteins From Two Dimensional Gels. *Proceedings of the National Academy of Sciences USA* 93: 14440-14445.
- [5] Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis T.; and Mann; M. (1996). Femtomole Sequencing of Proteins from Polyacrylamide Gels by Nano Electrospray Mass Spectrometry. *Nature* 379: 466-469.
- [6] Yaron, A.; Hatzubai, A.; Davis, M.; Lavon, I.; Amit, S.; Manning, A.; Andersen, J.; Mann, M.; Mercurio, F.; and Ben-Neriah, Y. (1998). Identification of the receptor component of the IkkappaBalpha-ubiquitin ligase. *Nature* 396: 590-4.

Exploiting Protein Structure in the Post-Genome Era

*Michael J. E. Sternberg*¹

*Paul A. Bates*¹, *Lawrence A. Kelley*¹, *Robert M. MacCallum*¹, *Arne Müller*², *Stephen Muggleton*², *Marcel Turcotte*¹

(1) Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, London, England <http://www.icnet.uk/bmm>. (2) Department of Computer Science, University of York, York, England

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.