Pacific Symposium on Biocomputing 7:398-399 (2002)

GENOME, PATHWAY AND INTERACTIONS BIOINFORMATICS

PETER KARP, PEDRO R. ROMERO
Bioinformatics Research Group, SRI International
333 Ravenswood Ave., Menlo Park, CA 94025, USA
{pkarp,promero }@ai.sri.com

ERIC NEUMANN
Beyond Genomics
40 Bear Hill Road, Waltham, MA
Eneumann@BeyondGenomics.com

The completion of major metazoan genomes, such as the human genome, has propelled life science research into a new era. Now that the "Book of Life", as some have called it, has been sequenced, and the vocabulary (genes) is being catalogued, the task now at hand is to identify the syntax and semantics of the book, and make sense out of what currently looks to us more like Lewis Carrol's "Jabberwocky" than Shakespeare's "Hamlet". The research and development for the next generation of bioinformatics tools for this task is on our critical path to unlocking the secrets of the human genome.

At the heart of this new challenge is the understanding of the interplay of genes and protein products. From various kinds of interactions (protein-gene, protein-protein), causal, regulated networks of biological pathways arise. Such networks are responsible for the development, maintenance, and responsiveness of all living systems. The collection and organization of pathway information is critical and still needs to be effectively addressed. Assimilating such information and turning it into knowledge of how living systems function (or function aberrantly in disease states) will become increasingly important for both basic research and drug discovery.

A key driving factor in our ability to understand how biological systems function is the emergence of new high-throughput functional-genomics technologies. Data from various kinds of experiments that have a bearing on pathways are being created at an increasing rate. The large ensemble of information they produce contains patterns that are a reflection of pathway dynamics, and therefore can be used to deduce pathway causal structures. These technologies and the information they generate include micro-array technologies that produce gene expression profiles, and 2-hybrid systems that produce information about protein-protein interactions.

Key to advancing our knowledge of biochemical pathways and networks is the intelligent analysis and mining of functional-genomics data in order to infer pathways and their regulation. For example, gene-expression profiles are dependent on the actual pathways that are in place within the target tissue, and can themselves be used to determine gene regulatory mechanisms as well as

signal transduction cascades. Expression data is a source of pathway "causal" information.

In order to help elucidate these functional relationships, researchers are applying a wide range of approaches to analyzing micro-array data. Methodologies such as Boolean networks, Bayesian networks, genetic algorithms, and simulation analyses are being used to help build new pathways or extend existing ones. These approaches each have their own advantages and disadvantages, and must be compared using a well defined set of expression benchmark problems. This session attempts to address the benchmark problem in particular. Only in this way will researchers be able to objectively evaluate different kinds of approaches for analyzing such diverse information.

Information collected from protein interaction sources (both experimental and literature-based) will also yield information about pathways, but from a physicalassociation perspective. Evidence from various kinds of protein-protein interaction experiments, such as 2-hybrid, hybrid-competitive, and 3-hybrid systems, will suggest the outlines of protein-binding cascades. Conclusive proof is not forthcoming because the partial protein domains that are created in such experimental systems may often give rise to false-positive and false-negative signals. It is from the careful analysis of such data, and its corroboration with other information, that the actual pathways will emerge.

Once pathways are elucidated, a new round of challenges present themselves. How do we compare and align individual pathways, and entire biochemical networks? Can we infer network properties when the immense number of quantitative parameters that govern their behavior is not known? Can we use estimation methods to infer the values of those parameters? How accurately can pathway behavior be simulated when parameter values are known?

Pathway bioinformatics still includes many challenges, and holds many promises for our basic understanding of biological systems, for drug discovery, and biotechnology.