Meeting report

From sequence to consequence

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham MA 02454-9110, USA.
E-mail: petsko@brandeis.edu

Published: 27 April 2000

Genome ***Biology*** 2000, **1(1)**:reports406.1

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/1/reports/406

A report from the 'Quantitative challenges in the post-genomic sequence era' workshop and symposium, San Diego, January 11-15, 2000.

Like drowning men, scientists awash in sequence data are grasping at any approach that can take them from primary structure to biological function: "from sequence to consequence", in Dagmar Ringe's memorable phrase. Although the derivation of function (however defined) from some level of structure was a recurring theme of the meeting, examples of approaches that might have some level of generality were scarce at this La Jolla 'interfaces in science' meeting.

The most widely applicable approach came from David Eisenberg (University of California at Los Angeles), who described two computational techniques for teasing some functional information out of sequence data. Both methods are easy to apply and can be carried out on a genome-wide scale. The first uses what Eisenberg refers to as phylogenetic profiling: determining which organisms possess homologs of a particular gene of unknown function. Other, nonhomologous proteins of known function that display the same phylogenetic profile are likely to function in the same pathway as the unknown protein. The approach is based on the assumption of correlated evolution, a phenomenon that arises because entire pathways tend to be conserved or deleted as species evolve. The limiting factors of such a strategy include the requirement for complete genome sequences across a wide evolutionary time scale (currently a problem because so few eukaryotic genomes have been sequenced, but this should change rapidly) and the obvious fact that housekeeping and very ancient pathways are likely to be found in nearly all organisms. Still, some success has been obtained already (Pellegrini *et al.*, *Proc Nat Acad Sci USA* 1999, 96:4285–4288). The big drawback would seem to be that the deduced function will frequently be so broad that it will be of limited utility.

This brings us to the second method, which relies on correlated domains in proteins, or what Eisenberg has termed the Rosetta Stone method. The rationale behind this method is the observation that some pairs of interacting proteins (in these studies, one partner in each pair will be of unknown function) have homologs in another protein that are fused into a single polypeptide chain. Although this fact has been known for decades – many bifunctional enzymes, such as chorismate mutase-prephenate dehydrogenase (involved in two consecutive reactions in tyrosine synthesis), and phosphoribosylanthranilate isomerase-idoleglycerolphosphate synthase (involved in two stages of tryptophan synthesis, are encoded by separate genes in other organisms – Eisenberg seems to be the first to suggest that it could be applied systematically to the decoding of function from sequence. In particular, he notes that many proteins are linked in this way to more than one other protein (for example, protein B may be fused to protein A in one organism and to protein C in another), helping to define parts of pathways or multiprotein complexes.

The real power, of course, comes from combining both methods, which is what he has now done (Marcotte *et al.*, *Nature* 1999, 402:83–86). He has shown that phylogenetic and Rosetta Stone linkages can be combined to suggest a general function (such as 'involved in translation') for more than half of the 2557 previously uncharacterized proteins in the genome of the budding yeast *Saccharomyces cerevisiae*. By similar analysis, functions for more than 40% of the 1521 previously uncharacterized proteins in the genome of *Mycobacterium tuberculosis* have also been assigned. Simple considerations, for example making the assumption that a gene of broad function is likely to be essential, permit 50 of the 3924 proteins in this pathogen to be chosen as possible drug targets. Such assumptions may not be universally true for functional genomics, but, nevertheless do succeed in providing consequence from sequences.