Computing protein function

Sarah A. Teichmann and Graeme Mitchison

We are more than the sum of our genes. Once genome projects have provided us with the list of an organism's genes, we have to determine how the proteins encoded by those genes cooperate to create and maintain the organism. This is a formidable task for the experimentalist, and even at the rapid pace of current large-scale biological projects the full picture may be slow to emerge. It turns, out, however, that there are some useful clues to be picked up by imaginative mining of existing databases. Two recent papers[1,2] have used patterns of co-occurrence to spot likely associations among gene products, and have come back with a rich haul of suggestive data. These data are likely to be valuable for many purposes, including the assignment of function to genes.

One of the methods draws on gene fusion. If two genes are separate in one organism and fused in another, then it is likely that the products of the two separate genes either belong to a protein complex or catalyze consecutive steps in a pathway. The protein made by the fused gene will consist of a single polypeptide chain in which the two protein subunits are covalently tethered, so they can transmit signals to each other or pass on substrates more efficiently. A researcher can recognize instances of gene fusion by scanning the genomes of many organisms and looking for situations where genes are separate in one organism and fused in another.

Enright and colleagues[2] found 64 fusion events from a survey of the genomes of two bacteria, one archaeon, and yeast. Almost all of the fused proteins of known function were enzymes. It is clear from the data that gene fusion among enzymes is as common in prokaryotes as eukaryotes: The yeast genome, with a genome of about 6,000 proteins, has 28 fused genes, whereas the three prokaryote genomes, which have about 7,500 proteins taken together, have 36. This is at first sight surprising, because eukaryotic proteins are longer on average than prokaryotic ones. Eukaryotic proteins therefore seem more likely to provide instances of fusions. Although it is true that eukaryotic genes are on average considerably longer, it is also the case that the bulk of sequences are only slightly longer in eukaryotes than in prokaryotes[3]. The principal difference is that eukaryotes contain more very long sequences, which constitute only a small fraction of the whole genome. These very long sequences may well be signal transduction proteins, cell surface receptors, and structural components such as muscle proteins.

Marcotte and colleagues[5] were able to detect a far greater number of pairs of genes in *E. coli* that are fused in another genome (6,809 in all). They achieved this by searching for fused proteins in two databases, Swissprot[6] and Prodom[7], rather than confining the search to a few other genomes. The greater number of hits may also reflect less stringent criteria for matching of proteins. In particular, they allow matches between subdomains of proteins. As some "promiscuous domains" are widespread components of many proteins, the fusions found this way may simply represent permutations and combinations of a set of common components and may not imply interactions. In fact, they find that, if the most frequently occurring domains are eliminated, the number of hits to Prodom falls dramatically from 3,531 to 749. Nonetheless, they do correctly predict many known interactions, including entire synthetic pathways. Interestingly, although they do not state what types of proteins they find, the particular examples they describe are all enzymes.

Pellegrini and colleagues[8] also consider a second method of detecting associations, which they call a "phylogenetic profile." This is based on the co-occurrence of proteins in different genomes. If two proteins are either simultaneously present or absent in all genomes, then it is likely that they are involved in the same functional pathway or grouping. The associations found by the method are likely to constitute a somewhat different set from those found through fusions, as the former are presumably genes that cooperate in some general, perhaps metabolic, way, whereas the latter are likely to correspond to physically tight-knit interactions. One might expect phylogenetic profiles to be inherently noisy, as the presence of a gene may be overlooked if it has diverged so far that a sequence comparison program cannot recognize it. Despite this caveat, eight times as many of the pairs determined by fusion were recognized by phylogenetic profiles compared to randomly chosen pairs.

In a still more recent paper, Marcotte and colleagues[1] added a third method of detecting associations, although this method is not purely computational but rests on the experimental determination of expression of genes under different conditions. Using DNA microarrays, it is possible to find those genes that are expressed at any one time as messenger RNA. Clustering genes based on the pattern of mRNA expression under different conditions, the expression profile reveals looser, less specific associations between proteins than phylogenetic profiles and gene fusions. Merging the results from all three methods, Marcotte et al. derive a rich web of associations with almost 100,000 links within the yeast genome. As these include the fusion links from "promiscuous domains," one might wonder how many of the links are to be trusted, and even if the domains in question are removed an uncertainty remains with all the methods.

Uncertainty is not new to bioinformatics. With sequence matching, for instance, assessing what is a reliable match is a subtle problem that requires a formal probabilistic treatment and careful testing against experimental data[9–11]. For protein–protein interactions, an example of such experimental data sets could be the recently completed genome-wide two-hybrid analysis of yeast[12]. Associations could also be given a probabilistic treatment, allowing different types of information to support or oppose each other. Perhaps something can be borrowed from automatic procedures for medical diagnosis, where graphical models have been developed to carry out automatic inferences from a variety of kinds of data[13]. For associations of proteins, the inference task is likely to become more interesting as the complexity of the network of associations increases, for then one can gauge the probability of links not only from the type and quality of data that supports them, but also from the larger graph of links in which they are embedded. The fact that the network found by Marcotte et al. already includes some examples of chains and circuits of links suggests that there is a rich scope for inference here, and, as genomic data accumulate, a fertile future for this field.

Sarah A. Teichmann (sat@mrc-lmb.cam.ac.uk) and Graeme Mitchison (gjm@mrc-lmb.cam.ac.uk) are scientific staff at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

1. Marcotte, E.M. et al. *Nature* **402**, 83–86 (1999).
2. Enright, A.J. et al. *Nature* **402**, 86–90 (1999).
3. Gerstein, M. *Fold. Des*. **33**, 518–534 (1998).
4. Bairoch, A. *Nucleic Acids Res*. **27**, 310–311 (1999).
5. Marcotte, E.M. et al. *Science* **285**, 751–753 (1999).
6. Bairoch, A. & Apweiler, R. *Nucleic Acids Res*. **27**, 49–54 (1999).
7. Corpet, F., Gouzy, J. & Kahn, D. *Nucleic Acids Res.* **27**, 263–267 (1999).
8. Pellegrini, M. et al. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
9. Altschul, S.F. et al. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
10. Krogh, A. et al. *J. Mol. Biol.* **235**, 1501–1531 (1994).
11. Park, J. et al. *J. Mol. Biol*. **284**, 1201–1210. (1998)
12. Uetz, P. et al. *Nature,* in press (2000).
13. Spiegelhalter, D.J. et al. *Stat. Sci.* **8**, 219–283 (1993).