Vol. 14 no. 5 1998
Page 383

Editorial

The Delphic boat or What the genomic texts tell us

The oracle of Delphi had the habit of questioning passers-by. One of the questions told the following story. I have a boat made of wooden planks. As time elapses they rot one after the other. At some time no original plank still remains in the boat: is it the same boat? Clearly the owner will say, yes. And he will be right. The boat is not the matter of the boat, but something else, much more interesting, that orders the matter of the planks: it is the relationships between the planks.

In a very similar way the study of life should never be restricted to the study of objects, but must study their relationships. This is why genomes cannot be considered simply as collections of genes. They are much more. How can we have access to this? Considering the current flow of genome sequences that are published, two contrasting images emerge: at first sight genes appear to be distributed randomly along the chromosome. In contrast their organisation into operons suggests that, at least locally, related functions are in physical proximity. In order to try to understand genome organisation, we must therefore explore the distribution of genes along the chromosome, but we should do this by generalising the concept of neighbourhood to many more types of vicinities than the mere succession of genes in the genomic text.

Our first observations suggest that this order is far from random, but is linked to the function of genes in relation with the cellular architecture. These results are fragmentary, so they must be experimentally validated. This ought to combine in silico analysis of the genome (bioinformatics) of model organisms, such as *E.coli* or *B.subtilis*, with their study *in vivo* (reverse genetics and physiological biochemistry, in particular using two-dimensional protein electrophoresis), and comparative studies with other genomes, with biochemical and structural analyses. If indeed the map of the cell is in the chromosome, this asks for some physical principle linking the succession of the genes — a symbolic text — and the cell's architecture — concrete matter. If we do not claim a divine principle, this should be a simple physical principle. The winning triplet of Darwinian natural selection (variation/selection/amplification) shows that evolution creates functions, that functions "capture" structures, so that structural analysis only becomes important when functions are understood.

The simplest way to evolve is to follow the arrow of time, to increase entropy. In water this is indeed the driving force for the construction of many a biological structure: this is at the root of the universal formation of helices, this allows the folding of proteins and the formation of viral capsids. But it should not escape our attention that the largest increase in entropy of a molecular complex in water occurs when the ratio surface/volume is the highest: when a planar structure is formed it orders the water molecules on both its faces. As a consequence, if this plane meets another one, it will lose one layer of water molecules, and stick there. Formation of planar layers should therefore be a very strong organising principle. Is it possible to find out, just knowing the genomic text whether a gene product will form such layers, whether it simply forms hexagons, for example? This is even more unlikely than that an amino acid sequence could tell us exactly the fold of a protein, without knowing pre-existing folds: pancreatic RNase would fold indeed, because selection isolated it for that (it is secreted in bile salts), but this would never be accepted as the paradigm of protein folding.

However, in silico analysis permits us to organise knowledge, and this might be a way to proceed in the future. In order to generate new knowledge, why not explore neighbourhoods of biological objects, considering genes as starting points, stressing that each object exists in relation to other objects. Inductive exploration will consist in finding all neighbours of each given gene. "Neighbour" has here the largest possible meaning. This is not simply a geometrical or structural notion. Each neighbourhood is meant to shed specific light on a gene, looking for its function as bringing together the objects of the neighbourhood. A natural neighbourhood is proximity on the chromosome: operons show that genes neighbours from each other can be functionally related. Another interesting neighbourhood is similarity between genes or gene products. The isoelectric point often gives a first idea of a gene product compartmentalisation. Also, a gene may have been studied by scientists in laboratories all over the world. And it can display features that refer to other genes: its neighbours will be the genes found together with it in the literature. Finally, there exists more complex neighbourhoods, the study of which gives particularly revealing results: two genes may be neighbours because they use the genetic code in the same way. One can also study all genes that belong to the same neighbourhood in the cloud of points describing codon usage of all the genes of the organism.

From the methodological standpoint this requires construction of neighbourhoods files (conveniently available to scientists in databases: a field of choice for bioinformatics). Finally, systematic investigation of bibliography will identify literature neighbourhoods, not only using title and abstracts, but the whole content of articles. We do not possess heuristics permitting direct access to unknown functions, and apart from preliminary studies there does not exist many places where such in silico work is developed. There exists however an excellent illustration of the concept of neighbourhood, the software Entrez, created by D.Lipman and colleagues at the NCBI.

All this has some flavour of a once fashionable field, Artificial Intelligence, but I do not wish to start that debate again! But this should also make clear to us that in silico analysis will never replace validation *in vivo* and *in vitro*: let us hope that propagation of erroneous assignments of functions by automatic interpretation of the genomic texts will not hinder discoveries. Knowing genome sequences is a marvellous feat, but it is the starting point, not the end.

Antoine Danchin