Developments in Industrial Microbiology - GMBIM
Edited by R.H. Baltz, G.D. Hegeman and P.L. Skatrud

Chapter 3

Comparative genome analysis: From sequences to functions

P. Bork

With the sequences of whole genomes in hand, the big challenge is to close the gap in knowledge between the genotype and phenotype of whole organisms. Bioinformatics is supposed to be a key player in this process. However, our current understanding of cellular processes is still very incomplete; thus a major part of the sequenced material cannot be appropriately interpreted.

The most powerful tools currently applied are database similarity searches with the hope to transfer the known function of a database protein to the unknown query. The large amount of data created by world-wide sequencing efforts or, more specifically, the task of annotating a whole genome by sequence analysis calls for automation in data handling and analysis. This requires accurate storage and updating mechanisms as well as appropriate retrieval software. While molecular databases are the most valuable source of information for comparative analysis, they are, like the accessing software, also the product of history and far from perfect. Thus, at present, working with sequence databases requires knowledge about their powers and their pitfalls (Bork and Bairoch, 1996; Bork 1996).

More than 80% of all known genes have at least one identifiable homologue in current databases; for the majority of them, functional predictions are possible (Fig. 1). However, the transfer of functional information is very difficult to quantify automatically, i.e., in the majority of cases only some functional features are shared by the query and in many cases a detailed manual analysis is required to avoid overinterpretations. Currently, automatic methods transfer the information of the best or the best annotated of the database hits. This usually leads to overpredictions of gene functions and hampers the reconstruction of nets of gene products and their pathways.

Even with a careful analysis, further automation of higher level predictions is difficult. The term "function" is loosely defined and computer-aided predictions can reveal protein features at different levels, such as molecular properties (e.g. cofactor binding site), pathway information (e.g. involvement in vitamin biosynthesis), cellular functions (e.g. role in transport mechanisms) or physiological aspects (e.g. essential for wing development). Furthermore, most of the proteins have several functions. Sometimes, only the dysfunction (disease genes) is known. Other functional features such as postranslational modification (proteins are often only the structural scaffold for a function that is realized via covalently attached carbohydrate moieties), expression patterns, tissue specificity, etc. are rarely considered in an automatic functional prediction.

Most of the known functional features in databases can be assigned to the molecular level. In particular, metabolic enzymes that constitute some 30% of the total amount of genes in bacteria are well-characterized. Thus, the next step in the understanding of cellular systems is to connect individual functions to pathways. That even basic

Peer Bork, EMBL, Meyerhofstr. 1, 69012 Heidelberg and Max-Delbrück-Centre for Molecular Medicine, Germany

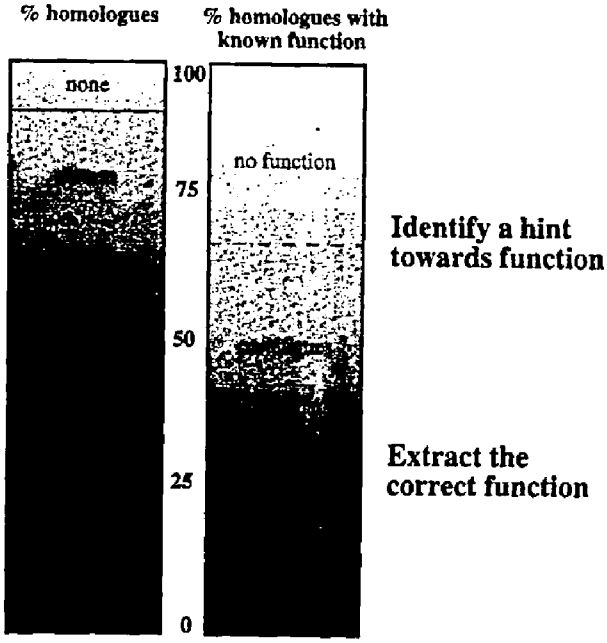

Fig. 1. The probability of a newly sequenced gene to have a homologue in current (1996) molecular databases (left column) and to be able to make some reliable function prediction (right column). This probability is based on several large scale sequence analysis projects (Bork et al., 1992; 1995; Casari et al., 1995; Tatusov et al., 1996). Note that these probabilities increase in time (Koonin et al., 1994) and that they represent an average of several species, i.e., they will be smaller for human genes and higher for bacterial ones. The dashed line in the left column gives a rough estimate as to the sensitivity of standard database searches such as blast (Altschul et al., 1994). Further sensitivity can be gained by investing the "twilight zone" of blast searches and applying motif and profile searches (Bork and Gibson, 1996). Functional predictions are difficult to quantify and there are no methods described to discriminate real orthologues (the corresponding gene in another organism) from paralogues (other homologous members of a multigene family). Thus, many genes in databases that have been annotated based on sequence similarities should be treated with caution (Bork and Bairoch, 1996).

pathways can vary in bacteria demonstrates the fact that in the first three complete sequence bacterial genomes, the citric acid cycle was incomplete or absent. Even if several genes coding for a known pathway are present in the complete gene pool of an organism, its reconstruction can be quite difficult. The enzymes might be divergent or they might be missing and this particular organism has either developed a variation of the known pathway or some steps are not needed as an intermediate is taken from external sources, as often can be observed for parasites. Furthermore, the undetected missing enzymes of a pathway might have been displaced by unrelated proteins that nevertheless carry the same function. This phenomenon, called "non-orthologous gene displacement" occurs, apparently quite frequently, as at least 16 out of 470 genes of *M. genitalium* fall into this category (Koonin et al., 1996b).

The availability of complete gene pools of model organisms also allows one to analyze their functional composition as well as to study the evolution of proteins and pathways. Classifying all identified gene products into several functional categories reveals a drift of protein function from metabolism to regulation and communication during evolution. More "modern" eukaryotic proteins, involved in communication and regulation, tend to have a modular architecture, i.e., consist of many structurally and often functionally independent building blocks that allow rapid evolution and that hint to multifunctionality (Bork et al., 1996).

In those modern proteins, more often than not only some domains can be identified and as these domains might function differently in different proteins, functional predictions are rarely possible. Nevertheless, even small domains in larger proteins might reveal their cellular localization or at least some functional features; important information if the task is to characterize, for example, positionally cloned disease genes.

An example is the breast cancer gene BRCA1 that has been sequenced in 1994 but no conclusion as for the function could be drawn (Miki et al., 1994). More recently there have been several reports on the cellular localization of BRCA1 that differed considerably placing it in the nucleus, in the cytosol or even in the extracellular environment (reviewed in Bork et al., 1997). Using sensitive motif and profile search techniques, we found a small duplicated domain of 100 residues at the C-terminus of BRCA1 to be related to a number of nuclear proteins (Koonin et al., 1996). Detailed inspection revealed that all the proteins containing this domain dubbed BRCT (Koonin et al., 1996) are involved in DNA damage-responsive cell cycle checkpoints (Bork et al., 1997). In the meantime, many experimental reports support this hypothesis; for example, it has been shown that the BRCT domain is a transcription factor (Chapman and Verma, 1996). Thus, the prediction not only provided arguments for a cellular location, but also offered a functional explanation for the role of BRCA1 (certainly, BRCA1 is a multifunctional molecule and many more studies are needed for the understanding of its molecular interactions and roles in cellular processes.

Taken together, sequence analysis is making a transition from a service performed after experimental characterization of a gene to a role in guiding further experiments based on predicted features of the gene product. Nevertheless, computational analysis of genomic sequences is only one set of methods, and it should be complemented with various experimental tools to advance our understanding of such complex systems as living cells.

References

Altschul, S.F., Gish, W., Boguski, M.S. and Wootton, J. (1994) *Nature Genet.* 6, 119-129.

Bork, P. (1996) *Science* 271, 431-432.

Bork, P. and Bairoch, A. (1996) *Trends Genet.* 12, 12, 425-427.

Bork, P. and Gibson, T. (1996) *Methods Enzymol.* 266, 162-184.

Bork, P., Downing, K.A., Kieffer, B. and Campbell, I.D. (1996) *Quart. Rev. Biophys.* 29, 119-167.

Bork, P., Hofmann, K., Bucher, P., Neuwald, A., Altschul, S.F. and Koonin, E.V. (1997) *FASEB J.* 11, 68-76.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) *Prot. Sci.* 1, 1677-1690.

Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W. and Gillevet, P.M. (1995) *Mol. Microbiol.* 16, 955-967.

Casari, G., Andrade, M.A., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, J., Valencia, A. and Sander, C. (1995) *Nature* 376, 647-648.

Chapman, M.S. and Verma, I.M. (1996) *Nature* 382, 678-679.

Koonin, E.V., Bork, P. and Sander, C. (1994) *EMBO J.* 13, 493-504.

Koonin, E.V., Altschul, S.F. and Bork, P. (1996) *Nat. Genet.* 13, 266-268.

Koonin, E.V., Mushegian, A. and Bork, P. (1996) *Trends Genet.* 12, 334-336.

Miki, Y. et al. (1994) *Science* 266, 66-71.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E. and Koonin, E.V. (1996) *Curr. Biol.* 6, 279-291.