Computational approaches for functional genomics

Kalin Vetsigian

October 31, 2001

The rapidly increasing number of completely sequenced genomes have stimulated the development of new methods for finding functional linkages between proteins [1]. A traditional bioinformatics method for extending the knowledge of protein function coming from biochemical and genetic experiments is sequence comparison. The idea is that genes with similar sequences in different species are most likely derived from a common ancestor gene which, in addition, has preserved its function during speciation because of selective pressure (once nature has found a good solution it will stick to it). Though this method is very powerful and gives some functional annotation for many of the newly sequenced genes, function inference based on homology is only one of many possibilities for analyzing genome data. In this paper I will briefly review techniques which provide information complementary to direct sequence comparison: phylogenetic patterns, gene fusion (Rosetta Stone) and chromosomal proximity.

The term gene/protein function can be understood at many different levels. The most basic level treats genes as stand-alone entities: what is the phenotype of a gene, what are the reactions a protein catalyzes. At higher level one tries to identify networks of relations which exist between different genes. Simple examples of such relations are multi-protein complexes (proteins physically interacting with each other) or proteins controlling different steps of a metabolic pathway. Such functional relations cannot be discovered by direct sequence comparison because proteins in the same protein complex are not necessarily homologous, nor are proteins at different stages of the same pathway.

Phylogenetic pattern methods study the correlated evolution of non-homologous genes. One takes a gene and tabulates the presence or absence of orthologs of it in other genomes. (Notice that knowledge of complete genomes is required in order to ascertain that no ortholog is present.) Then one compares phylogenetic profiles of different genes and searches for patterns. A striking example is given in Figure 1 (taken from [1]). Two genes with completely different sequences have almost perfectly complementary phylogenetic profiles. Such strong correlation cannot be attributed to chance. Probably the best explanation is non-orthologous gene displacement (NOGD), i.e at some point in the evolution one of the genes functionally replaced the other. Certainly NOGD tends to produce complementary evolution patterns because if at certain point we have

Table 1. Complementary phylogenetic patterns, non-orthologous displacement and prediction of protein functions

	Species[a]																	
	Archaea				Eukaryota		Bacteria											
Pathway/ Enzyme	Af	Mj	Mth	Ph	Sc	Aa	Tm	Ssp	Ec	Bs	Mt	Hi	Hp	Mgp	Bb	Tp	Ctp	Rp
Translation[b]																		
Class II lysyl-tRNA synthetase (COG1190)	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	+	+	-
Class I lysyl-tRNA synthetase (COG1384)	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+
Glycolysis[c]																		
FBA (COG0191)	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-
DhnA-type FBA (COG1830)	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	-	+	-
Thymidylate[d] **biosynthesis**																		
Thymidylate synthase (COG0207)	+	+	+	-	+	-	-	-	+	+	+	+	-	+	-	-	-	-
Predicted novel thymidylate synthase (COG1531)	-	-	-	+	-	+	+	+	-	-	+	-	+	-	+	+	+	+

[a]Aa, *Aquifex aeolicus*, Af, *Archaeoglobus fulgidus*, Bb, *Borrelia burgdorferi*, Bs, *Bacillus subtilis*, Ctp, *Chlamydia trachomatis & pneumoniae*, Ec, *Escherichia coli*, Hi, *Haemophilus influenzae*, Hp, *Helicobacter pylori*, Mgp, *Mycoplasma genitalium & pneumoniae*, Mj, *Methanococcus jannaschii*, Mth, *Methanobacterium thermoautotrophicum*, Ph, *Pyrococcus horikoshii*, Rpr, *Rickettsia prowazekii*, Sce, *Saccharomyces cerevisiae*, Ssp, *Synechocystis sp.*, Tm, *Thermotoga maritima*, Tp, *Treponema pallidum*. [b]In Tp, the only functional lysyl-tRNA synthetase is probably the class I enzyme; the class II enzyme is a distinct truncated form that is likely to have a function other than translation. [c]Aa and Ec possess both types of FBA; Rp lacks glycolysis. [d]Mt is predicted to posses both types of thimydylate synthase

Figure 1:

two genes in a genome performing the same function the selective pressure is relaxed and one of the two genes will start to mutate rapidly until it is completely unrecognizable by sequence comparison. Even for ancient NOGD the complementarity might not be perfect because the diverging gene can obtain a new function before its sequence have diverged enough. In certain cases, even a perfect complementarity might not imply NOGD. The first two profiles bring up the point because, roughly speaking, one of the genes is present only in archaea and the other only in eukaryotes and prokaryotes. There might be many other entirely unrelated genes which share the same pattern (present only in archaea for example). What makes the NOGD hypothesis somewhat convincing in this case is that in Bp and Rp bacteria species the pattern is reversed but still complementary. For the last two profiles on Figure 1 no such doubt arises.

Another obvious possibility is to look for similar profiles rather than complementary ones. Groups of functionally related genes are more likely to have similar phylogenetic profiles than unrelated genes. Two proteins that form a dimer, for example will tend to have similar profiles because if only one of them is inherited, the inherited one will loose its function and often diverge beyond

recognition. Besides predicting protein function, unexpected phylogenetic patterns, such as presence of orthologs in all but one bacteria species, sometimes help to identify genes that were omitted in genome annotations and were recognized only after a closer look at the raw DNA sequence data. This are usually small genes.

The fusion analyses (Rosetta Stone) methods are based on the observation that there are many proteins or protein domains that are separate in some species but their orthologs in other species are fused into a single gene. Functional link between the two genes is inferred. If two proteins catalyze consecutive steps in a metabolic pathway their fusion into a single protein will greatly increases the effective concentrations of the two enzymes - the product of the first reaction can immediately undergo the next step in the pathway because the enzyme needed is already there. Studies in E. coli have been shown that 75% of all fusion links indeed relate two metabolic genes [4]. An example of fusion is given on Figure 2 (taken from [5]). Fusion of proteins involved in protein complexes is also encouraged by evolution - different subunits don't have to find each other in order to form a complex. As an example, A and B subunits of type II topoisomerases are separate genes in bacteria but a single gene in eukaryotes. Another advantage valid for both complexes and pathways is that gene regulation is facilitated. The authors of the fusion methods Marcotte et al. [6] proposed that the opposite might also be true - two noninteracting proteins could evolve a strong affinity after being fused, and perhaps after subsequent separation they can become interacting proteins.

Functionally related genes in prokaryotes tend to form operons. If certain genes form operons in some lineages but not in others this would give indications that the corresponding genes are functionally related (in all genomes). Exact prediction of unknown operons is difficult but one can still detect groups of genes which are close on the chromosome in some genomes and scattered in others and consider that as an evidence of functional linkage.

All of the above methods rely on knowledge of "corresponding genes" in different genomes, i.e. genes which are likely to have a common function and are derived from a common ancestor. Usually a corresponding gene is identified with an ortholog. However, the situation is complicated by the presence of gene duplication events. On Figure 3 (taken from [7]) 1A is formally orthologous to all other genes. So what is the corresponding gene of 1A in species 4? It is reasonable to say that in each gene duplication event one of the genes retains the original function and the other diverges more rapidly in terms of sequence. Therefore the "corresponding gene" is usually the gene with greatest sequence similarity. We can accept this as a practical definition but we must also address the question of how we decide if a given gene has a corresponding gene in a given genome or not. This is crucial for the phylogenetic pattern method to work. One possibility is to set a threshold of statistically significant sequence similarity but this threshold should dependent on the time elapsed after the species in question diverged as well as on the stochastic model. In the most popular COG database [3] an attempt is made to avoid this ambiguity by constructing graphs of BeTs. The nodes of the graph are the proteins in all genomes. For each protein and

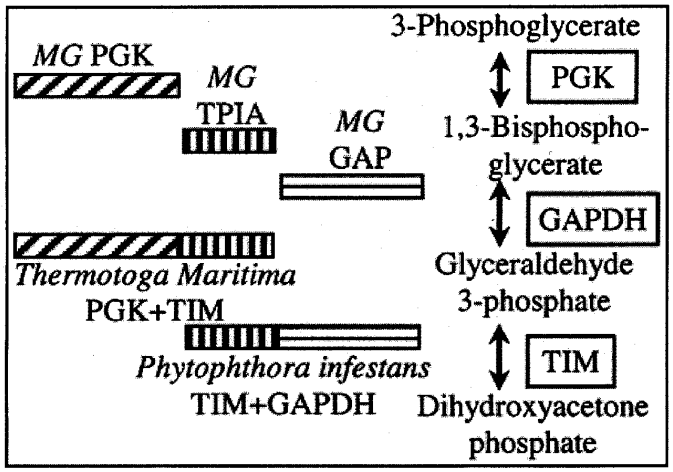

Figure 2: Correspondence between functional associations and genes linked by the fusion method. Independent genes in one genome may be found as one continuous gene in other genomes. These fusion links can confirm known functional relationships between genes

each genome there is a directed link, BeT, from this protein to the best matched protein in the genome. The objective of the algorithm is to recover clusters of orthologous groups (COGs). Each COG is defined as a set of proteins such that any two proteins from different lineages are orthologous. By this definition the COGs in Fig.3. are {1A, 2A, 3A, 4A}, {1A, 2B, 3B, 4B} and {1A, 2B, 3C, 4C, 4D}. The triangles formed in the graphs of BeTs are considered to form a COG. In addition triangles with common sides are merged into a bigger COG. For the distance tree shown on Figure 3d. the above algorithm correctly retrieves the COGs. Now suppose that 1A was deleted for some reason. Then the BeTs of 2A, 3A and 4A in species 1 are likely to be different due to the low overall scores of the BeTs and the finite differences between 2A, 3A and 4A. If this is so, the algorithm will correctly identify the absence of an ortholog of these genes in species 1. Once we have the database of COGs we may use it to build phylogenetic trees or to check if a given gene has a corresponding gene in a particular genome. For example, if we are constructing the phylogenetic profile for 2B we will know that 1A is the only ortholog in species 1. Then we can immediately see that 1A and 2A are also orthologs and conclude that 2A and 2B are paralogs. If the similarity between 2A and 1A is much greater that that between 2B to 1A then it is likely that the function of 2B have started to diverge from that of 1A and 2A. In this case we might choose to say that 2B has no corresponding gene is species 1.

At the end I would like to mention that there are other high-throughput experimental methods (besides genome sequencing) for generating massive amount

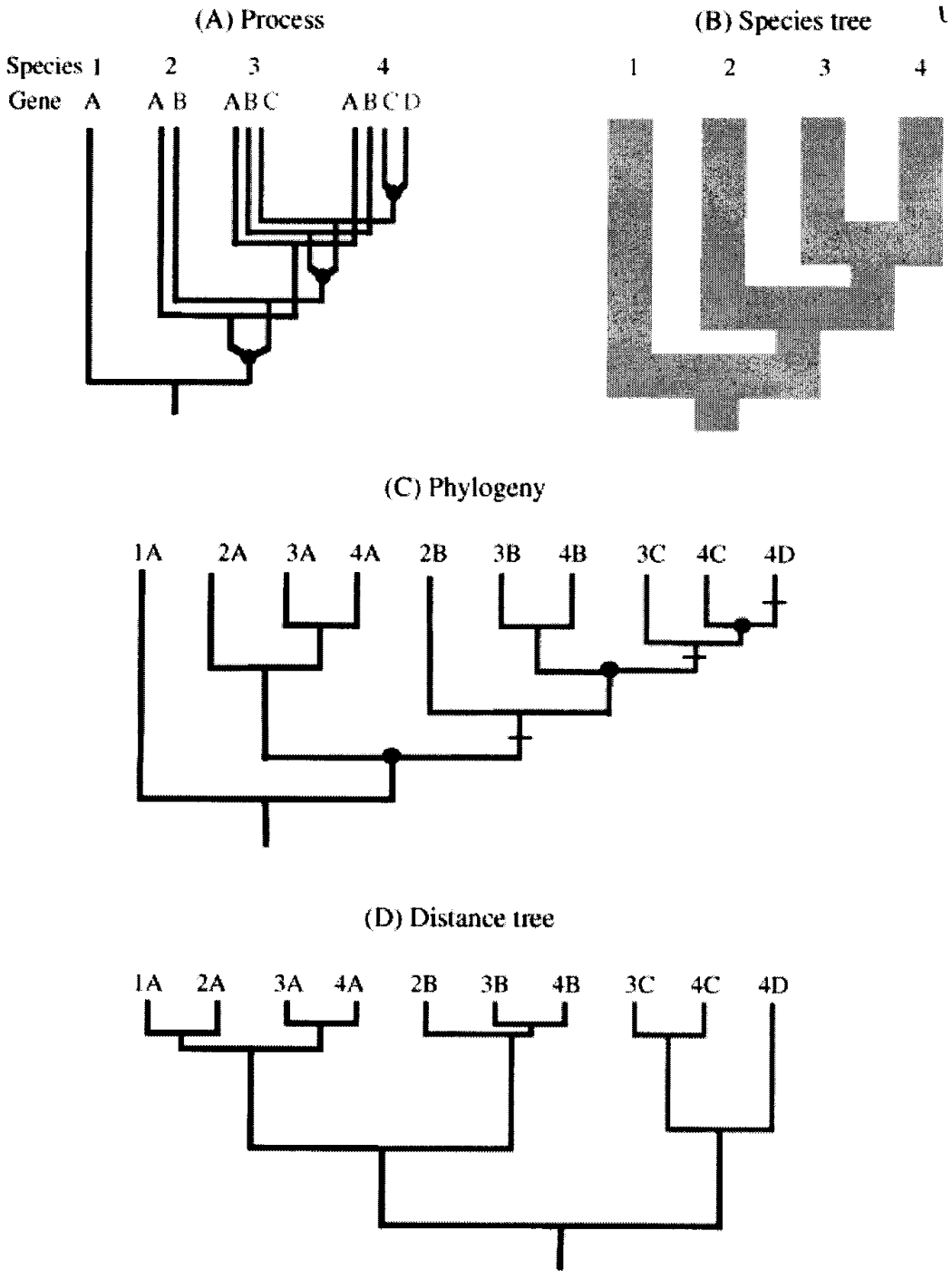

Figure 3: An example of gene family phylogeny. Gene duplication events are marked with filled circles. Speciation events are unmarked.

of data on gene interaction. DNA microarrays can measure the gene expression profiles of many (thousands) genes simultaneously in many different conditions. In addition, large-scale effort to measure directly protein-protein interactions is underway using several different techniques. Even though each of these methods doesn't give accurate enough predictions of meaningful protein-protein interactions combining and analyzing data from all these sources simultaneously can yield accurate and comprehensive maps of the interaction networks. This will make it possible to think at yet higher level about the biological complexity.

References

[1] Galperin, M. & Koonin E., *Nature Biotechnology* **18**, 609-613 (2000)

[2] Eisenberg, D., Marcotte, E., Xenarios, I. & Yeates, T., *Nature* **405**, 823-826 (2000)

[3] Tatusov, L., Koonin, E. & Lipman, D., *Science* **278**, 631-637 (1997)

[4] Tsoka, S. & Ouzounis, C. A. *Nat. Genet.* **26**, 141-142 (2000)

[5] Yanai, I, Derti, A. & DeLisi, C.*Proc. Natl. Acad. Sci.* **98**,7940-45, (2001)

[6] Marcotte, E., Pellegrini, M., Ng, H. L., Rice, D., Yeates, T. & Eisenberg, D., *Science* **285**, 751-753, (1999)

[7] Thornton, J. & DeSalle, R., *Annu. Rev. Genomics Hum. Genet.* **2000.01**, 41-73, (2000)