

# **Beyond Sequence Similarity, or Sequence Analysis in the Age of the Genome**

**Itai Yanai, Adnan Derti and Charles DeLisi**

**Graduate Program in Bioinformatics and  
Department of Biomedical Engineering  
Boston University  
Boston, Mass.**

Until recently, the only computational method available for predicting the function of an uncharacterized gene was sequence similarity, an approach that is effective but restricted to instances in which the function of a closely related sequence is known. Because homologs of characterized function are not available for many sequenced genes, new methods that do not rely upon sequence similarity are critical if we are to exploit and address the avalanche of sequences. Such methods have in fact emerged recently. By considering the genome as a parts list, we can link two genes functionally if they share the same evolutionary pattern, such that they are either both present or absent in any of the known genomes. Insofar as the genome is a permutation of genes, two genes may be associated if they are consistently found as chromosomal neighbors across genomes. Genes may also be linked if they are found fused as one gene in another genome or if they have common regulatory elements and/or similar expression patterns. Together, these methods constitute the many diverse senses of sequence analysis, which may collectively hint at the function of an uncharacterized sequence in the context of all other known sequences.

Keywords: non-homology-based function prediction algorithms, genomic context, comparative genomics, functional links, phylogenetic profiling, chromosomal proximity, domain fusion analysis, regulatory sequences, mRNA expression, gene chips, and DNA microarrays.

## Introduction

The transcription of a genome from organism to computer file, while a substantial technical feat in itself, offers its greatest contribution as a starting point for the study of newly sequenced genes of unknown function. Early approaches to gene identification started with an observed phenotype, usually a biochemical defect, and required years of study before the protein and then the gene were identified. The introduction of positional cloning in the 1980s accelerated the process by better than an order of magnitude and resulted, in short order, in the discovery of genes related to Duchenne muscular dystrophy, cystic fibrosis, retinoblastoma, fragile X syndrome and many other diseases. More recently, the ability to rapidly and inexpensively sequence complete genomes and to monitor the expression of thousands of genes simultaneously places us at the threshold of another acceleration in functional discovery, one that will be both quantitative and qualitative, and which will be made possible by the development of powerful computational methods.

### Sequence analysis by sequence similarity

Sequence similarity has been singularly successful in revealing clues to the function of newly sequenced genes. One of the most dramatic early discoveries was the similarity between the sequences of the transforming protein of simian sarcoma virus and human platelet derived growth factor, revealed by a database search using standard sequence alignment software (Chiu IM 1984). Another example was the assignment of the cystic fibrosis gene to a class of membrane transport proteins, from sequence alignment searches alone, without doing a single experiment. These searches are now conducted routinely in thousands of laboratories throughout the world. They supplement experimental information by using knowledge derived from one system to infer the function of a closely related gene sequence in another system.

### Sequence analysis *without* sequence similarity

As important as sequence similarity searches have been, the need for new methods not tied to homology is acute, primarily because the explosive growth in sequence information has revealed large numbers of genes that are not similar to any genes of known function. However, this avalanche, which includes the complete genome sequences of over thirty prokaryotes in addition to yeast, a worm, the fruit fly, and humans, has also provided us with a resource: we can now look at an unknown gene in its context in all the genomes where it appears.

In this chapter, we review, and to some extent critique, emerging context-dependent methods for revealing clues about the functions of genes (see Figures 1 and 2). These methods all rest upon the notion that genes do not function in isolation but are instead linked to one another (Danchin 1998); in other words, function is an attribute of a system, rather than a single component. The key to deciphering function is an analysis of these relationships (Eisenberg, Marcotte et al. 2000; Huynen, Snel et al. 2000).

The first method we review is phylogenetic profiling, which rests upon the notion that if the presence of two genes across the phyla is correlated, the genes are functionally related. If more than two genes are considered, we can begin to build functional networks. The key to this method is developing a sound, predictive relation between correlated phylogenetic profiles and function. Other methods are based on physical proximity in the genome: one might expect that when the short distance between two genes is conserved across phyla, that conservation is the result of a functional requirement. Similarly, genes that are distinct in a given organism but are found fused in another organism tend to be linked functionally. Finally, evidence is emerging that genes sharing similar regulatory sequences and/or correlated mRNA expression patterns tend to subserve a common function or set of functions.

Computational analyses based on these ideas result in predictions that suggest *in vivo* or *in vitro* experiments. The computer is thus essentially acting as a hypothesis machine, driving the experiments in the wet lab. As today's biology finds itself flooded with sequences, a most useful strategy is to filter the raw sequences through these computational methods to make a first pass at the data and point to interesting avenues of research.

## Phylogenetic Profiling

One of the generalizations emerging from the comparative study of genomes is that approximately 70% of the genes in a given genome are not distinct to that genome but can be found even in distant phylogenetic clades (Koonin, Mushegian et al. 1997). Given such a high recurrence of genes among the genomes, we discuss in this section how information about a gene can be derived from its distribution across the genomes.

### Clusters of Orthologous Groups

The Clusters of Orthologous Groups (COGs) system is an elegant method for examining the orthologous relations among the genes of the sequenced genomes (Tatusov, Galperin et al. 2000). A COG begins as a set of three genes from three distinct genomes where each pair, for example, gene a from genome A and gene b from genome B, is such that the gene most similar in sequence to gene a in genome B is b and vice versa. Two COGs are united to form one larger COG when both COGs include two of the same members. As such clustering steps are repeated, the COGs are enlarged until each one represents a complete set of orthologous genes. A particularly nice feature of this method is that since it looks for only the best match, it is not dependent on any threshold for significant sequence similarity and detects both quickly and slowly evolving sets of orthologs.

Once COGs are created, we can consider their representation among the sequenced genomes. We do so by constructing a phylogenetic pattern or profile (here, the two terms will be used interchangeably) which is simply an expression of the presence or absence in each genome of at least one member of the COG. For example, if the COG has at least one gene in all seven genomes in the initial study (Tatusov, Galperin et al. 2000), we assign to it the pattern “ehgpcmy”, where e, h, g, p, c, m, and y stand for *E. coli*, *H. influenzae*, *M. genitalium*, *M. pneumoniae*, *Synechocystis sp.*, *M. jannaschii*, and *S. cerevisiae*. Indeed, the pattern “ehgpcmy” was one of the two most abundant in the initial COG study (Tatusov, Koonin et al. 1997). The other was “eh-cmy”, i.e., present in all the genomes except *M. genitalium* and *M. pneumoniae*, the two mycoplasmas. Together, these two patterns account for only a third of the COG patterns, signifying that there is considerable variation in the patterns.

Can these COG patterns provide functional information about genes and genomes? The remainder of this section attempts to answer this question. Before proceeding, however, it is important to note that the construction of a COG is only one method of building a phylogenetic pattern. Given the gene's sequence, we build the pattern by simply querying, by sequence similarity, if that sequence is present or absent in each genome using an ad hoc threshold for similarity.

### Two ways to view the phylogenetic patterns

Given this concept of a phylogenetic pattern, we arrive at two distinct ways of using a list of such patterns. We can link two genes if they have similar patterns, i.e., co-occur, or link two genomes because they have the same set of genes. We can construct a matrix

where columns correspond to genomes and rows correspond to genes. An element of the matrix contains a 1 if that gene is present in that genome or a 0 if it is absent. Each row thus contains the phylogenetic pattern of a gene, while each column lists the set of genes in that genome. We can then look for similar rows (genes with similar phylogenetic profiles) or similar columns (genomes with similar genes).

- *Phylogeny based upon gene content*

By comparing the columns of the matrix, we are asking, “given that a genome is a *set* of genes from a library of genes (all the rows), how do the different sets/genomes compare?” If we assume that the more similar the gene contents of two genomes, the more recent their time of divergence, we can use a measure of similarity of their gene contents to construct evolutionary trees. One possible measure is the number of genes common to both genomes divided by the size, in genes, of the smallest genome (Snel, Bork et al. 1999). A phylogenetic tree derived from such a metric looks remarkably similar to the r16 RNA trees (Snel, Bork et al. 1999; Tekaiia, Lazcano et al. 1999).

One can also ask if there are genes or whole classes of genes that are kingdom-specific. For example, one study (Ouzounis and Kyrpides 1996) considered evidence that archaeal genes for transcription are significantly more similar to those of eukaryotes than to those of bacteria; the authors used this observation to make hypotheses about the presence of various cellular processes in the last universal ancestor.

Another benefit of comparing gene content among genomes is the possibility of identifying phenotypic characteristics of one genome by differentiating its gene content with that of another closely related genome that lacks the phenotype. Huynen et al. (Huynen, Dandekar et al. 1998) introduced this strategy and applied it to the *H. pylori* genome. By considering all *H. pylori* genes that do not have orthologs in *E. coli* and *H. influenzae*, the authors identified 123 genes which were known to function in host interactions such as membrane proteins. These genes can be investigated further concerning their putative role in *H. pylori*'s parasitic phenotype.

- *Functional coupling based upon phylogenetic patterns*

The orthogonal view of the matrix, based on clustering the phylogenetic patterns according to similar distributions in the genomes, is most relevant to this chapter. The fundamental hypothesis is that non-homologous genes whose patterns of occurrence across a representative set of genomes are highly correlated subserve a common function or set of functions.

#### Phylogenetic patterns of glycolytic enzymes

A number of studies have focused on specific associations among proteins and have clustered them by using a library of genomes (Huynen, Dandekar et al. 1999), (Dandekar, Schuster et al. 1999). To consider a specific set of genes using phylogenetic patterns, we examine the study of the glycolytic, Entner-Doudoroff, and pyruvate pathways as they

occur in 17 genomes (Dandekar et al.; see Figure 3). A number of interesting observations arise from this data. Looking at specific sections of the pathway, for example, pyruvate synthase, pyruvate dehydrogenase, and the lower part of glycolysis, the authors find that these phylogenetic patterns are fairly strongly conserved. Thus, relaxing the criteria that the patterns must match exactly, and allowing them to deviate by a few elements, allows more genes to be linked.

It is reassuring that the partitions assigned to the pathways (for example, Entner-Doudoroff or lower part of glycolysis) also roughly correspond to their evolutionary patterns. Dandekar et al. also show that the phylogenetic patterns of proteins that form complexes are much more conserved than those of proteins which do not form complexes (see pyruvate synthase and pyruvate dehydrogenase). In addition, we point out that when a protein is present in all the available genomes, as in the genes involved in the lower part of glycolysis, its phylogenetic profile is less informative, because there are likely to be many other genes present in all genomes (Tatusov, Koonin et al. 1997). For example, ribosomal proteins and DNA polymerases can be expected in all organisms and thus will have identical phylogenetic profiles, yet they are not directly involved in the same function or process.

On the other hand, this intersection of all genomes should help identify the common genes required for life (Mushegian and Koonin 1996). As more genomes become available, there is an increasingly smaller chance that a gene will be found in all of them; each additional genome doubles the number of phylogenetic profiles possible and hence reduces the probability that different gene families will exhibit the same pattern. At this exponential rate, the different patterns of evolution are expected to be resolvable to the point where each functional family will have its very own pattern.

#### Phylogenetic profiling as a predictive tool

Since a cluster of genes with identical phylogenetic profiles may consist of both genes of unknown function and others that have been studied experimentally, this method can be used to predict the functions by extending the known functions to the uncharacterized genes.

This concept was introduced by Huynen et al. (Huynen and Bork 1998), but its large-scale implementation had to wait until enough genomes were available to make the method feasible (Gaasterland and Ragan 1998; Pellegrini, Marcotte et al. 1999). For example, Pellegrini et. al found in *E. coli* only four genes with the same profile as ribosomal gene RL7, all of which had been previously studied and determined to be associated with protein synthesis. To generate more links between genes according to these profiles, the criteria for linking profiles was relaxed from "exact" to "similar", such that two genes could be linked even if their profiles differ at only one entry. (We describe below a more rigorous method to determine the significance of similar profiles.) With this slightly more permissive association as well as 16 genomes, the authors found 27 new proteins in this RL7 cluster. Of these, roughly half were annotated as ribosomal proteins while many others were hypothetical and/or uncharacterized proteins. One such

uncharacterized protein, SmpB, was recently confirmed to be involved in protein synthesis (Karzai, Susskind et al. 1999) (Marcotte 2000).

In addition to verifying that genes with the same phylogenetic profile have the same function, Pellegrini et al. tested the converse: do genes with the same function have the same phylogenetic profile? By grouping genes according to common keywords, they compared the profiles of clustered genes. Even with a looser requirement allowing up to two deviations among the profiles, relatively few of all the possible links among the functionally related genes were made using the profiles, though still appreciably more than random. An explanation for this is that the annotated function is too broad, uniting multiple phylogenetic patterns. Perhaps this method will be able to aid in annotating at the level of protein families, by classifying genes into families based on correlated evolution.

### Scoring correlations among phylogenetic profiles

In order to evaluate the significance of correlations among profiles, it is helpful to derive the probabilities of such occurrences. Suppose we have access to  $N$  fully sequenced genomes. Gene A occurs in  $k \leq N$  genomes; gene B occurs in  $r$  genomes in common with gene A as well  $s$  genomes not in common with gene A.

The probability of the observation depends on what we take as prior knowledge. To be specific, suppose we begin with the observation as stated above. Then, the number of ways in which gene B can co-occur  $r$  times with  $k$  occurrences of gene A is

$$\frac{k(k-1)\dots(k-r+1)}{r!} = \binom{k}{r}$$

The number of ways the remaining  $s$  B genes can be distributed over the remaining  $N-k$  genomes is

$$\frac{(N-k)(N-k-1)\dots(N-k-s+1)}{s!} = \binom{N-k}{s}$$

The probability of the particular observation is

$$2^{-N} \binom{k}{r} \binom{N-k}{s}$$

The probability that at least  $r$  of the B genes co-occur with the A genes (with  $s+r \geq k$ ) is

$$2^{-N} \sum_{j=r}^k \binom{k}{j} \binom{N-k}{s+r-j}$$

assuming all genomes are equally spaced evolutionarily.

If the phylogenetic distances between the genomes are not identical, then the expressions must be modified so that lower weights are assigned to associations arising from those genomes that are most closely related. The simplest procedure would be to use weights that are proportional to distances on the 16S rRNA tree.

The derivation above assumes that all phylogenetic profiles are equiprobable, however in reality profiles are not distributed evenly along profile space (see for example (Tatusov, Koonin et al. 1997)). To correct for this, the probability of finding each profile should relate its frequency in observed profiles.

The above remarks implicitly assume that the “genes” are composed of single domains, where a domain is an independently heritable unit. In the more realistic situation in which genes consist of more than one domain, reasonable probability estimates can still be made, provided that the domains boundaries of the genes of interest are known.

#### Summary: evolutionary considerations

Since phylogenetic profiling is based on the detection of co-evolved genes, it is interesting to consider under which conditions it will fail. For example, the correlation between two genes will be reduced if one or both of them is replaced by analogous genes in one or more genomes. Such non-orthologous gene displacement has been discussed by Mushegian and Koonin ((Mushegian and Koonin 1996)). For example, a class I lysyl-tRNA can replace the non-homologous but analogous class II gene in function, leaving each genome able to use either of the two but never having a need for both (Galperin and Koonin 2000).

Another, perhaps more pervasive, reason for errors in linking by phylogenetic profiling is the coupled processes of gene duplication and gene loss (Huynen and Bork 1998). Gene A may be duplicated to produce gene B which may then speciate to assume a function distinct from gene A. If gene A is subsequently lost from the genome, then gene B may be mistakenly assumed to be an ortholog of gene C, which was originally orthologous with A. Such errors are due to the difficulty in assigning orthologs.

More information can be expressed in a phylogenetic profile if consideration is given to the number of copies (paralogs) of the gene found in each genome. That is, in place of Boolean bits, a profile may be composed of integers, where each denotes the family size of the gene. A correlation between pairs of profiles may then be strengthened if the two correspond not only in their presence in similar genomes but also in terms of family sizes (Wu et al., work in progress).

As more genomes are being sequenced, we are obtaining an increasingly better sampling of “bags of genes”, which is how the phylogenetic profiling method views a genome. When a sufficient number of genomes are available, it will be possible to give a statistical score to each match among two profiles relating the significance of the correlation. An obstacle toward this goal is that an even phylogenetic distribution of all genomes will not be present, i.e., the available genome sequences do not represent an even sampling of life and instead focus on certain species. In some cases, different strains of the same species are sequenced (see *H. pylori* and *M. tuberculosis*). Since these strains have very similar or nearly identical gene contents, their inclusion in a phylogenetic pattern would bias the results in favor of the genes present in the over-represented genomes. Methods need to be developed to ensure that such a bias does not occur, for example by assigning different weights to the genomes.

## Chromosomal Proximity

A popular aphorism regarding the value of real estate advises that the three most important factors are: “location, location, location.” While it is an overstatement to suggest that the chromosomal neighborhood of a gene is crucial to its function, useful information about a gene can be derived from knowledge of the functions of physically proximate genes.

### Gene order along the chromosome

Knowing the entire genome sequence of an organism allows a compilation not only of a list of the gene sequences but also of a list their order along the chromosome. An immediate question is thus, “how much functional information is encoded in the particular permutation of genes observed in the genome?” A genome of  $N$  genes can be ordered in  $N!$  different ways, an enormous number. From a comparison of related genomes, we know that global order is not conserved, and that the distance between two genes in one genome is more likely to be conserved in another genome if it is short (Watanabe, Mori et al. 1997). Selective biases operate at the local scale of gene clusters, preserving the proximity of certain genes. We can thus view a genome as a more or less random permutation of gene clusters, i.e., the genes within a cluster are conserved but the order of the clusters is not. Furthermore, these evolutionary pressures tend to group genes of *common function*. Thus, in deciphering the function of a gene, we find clues by considering the functions of the adjacent genes in all the genomes where the gene is present.

### Two models for the origin of gene clusters

The notion that the proximity of certain genes is conserved across genomes is most interesting from the point of view of this chapter because of the functional associations among the genes. It is thus appropriate to consider what evolutionary pressures may act to cluster genes of common function. There are two dominant models to account for this effect: the “co-regulation model” (Pardee 1959) and the “selfish operon” model (Lawrence 1997). The former benefits the organism, while the latter benefits the gene cluster itself. By having the genes in proximity, the individual gets the selectable benefit of a facilitated regulatory mechanism, or so the “co-regulation model” reasons. The “selfish operon” model, on the other hand, infers that what motivates the formation of a gene cluster is the facilitation it confers to the horizontal transfer of that cluster. In this view, genes that can collectively impart a novel phenotype (metabolic, for example) cluster not for the benefit of the organism but for their own advantage of making them better suited for transfer. In other words, their organization as a cluster facilitates their transfer to other organisms.

### Restriction systems: an example of functionally-associated proximate genes

As a defensive measure against bacterial viruses (bacteriophages), bacteria have evolved a mechanism for destroying foreign DNA in the cell. A cell can carry a battery of the so-

called restriction enzymes that recognize specific short sequences in the DNA molecules and cleave at these sites. In order to avoid destroying its own genome, the cell also carries a corresponding set of methyl transferases that methylate the same sequences in the host DNA, thereby protecting the host from its own restriction enzymes. Notwithstanding the underrepresentation of palindromic words corresponding to restriction sites (Koonin), and the presence of methyl transferases that facilitate the correction of replication errors, the cell does not benefit from having an extra methyl transferase gene without the corresponding restriction enzyme, and suffers a terrible fate if it possesses a restriction enzyme without the corresponding methyl transferase.

Most conspicuously, in all known occurrences of a restriction enzyme gene, a methyl transferase gene is found adjacent along the chromosome (Richard Roberts and Richard Morgan, personal communication). Based upon the two models cited above, there are two interpretations for this phenomenon. According to the “co-regulation” model, both genes need to be turned on at about the same time in order for the defense system to work properly, and the proximity facilitates this co-regulation. In the case of restriction enzymes, however, the evidence weighs heavily against this interpretation, since the two genes, though adjacent, do not share a common regulatory scheme; the methyl transferase must and does start to protect the host DNA before the restriction enzyme launches its attack against any unprotected DNA. This implies that the alternate explanation is more valid, namely that the clustering of these two complementary genes facilitates the transfer of the genes, which together give the phenotype of protection against phages.

The proximity of these two genes is very useful in identifying restriction enzymes. Restriction enzymes, which are sold commercially for cloning, form an amazingly unconserved gene family and cannot be identified through sequence comparisons with other restriction enzymes. By contrast, methyl transferases have a motif that is strongly conserved within their gene family. Thus, the common technique used in identifying restriction enzymes is to find the methyl transferase by sequence analysis and look at the adjacent genes, one of which is most likely a restriction enzyme

#### Neighboring genes tend to be functionally related

Are there functional associations among adjacent genes? To rephrase the question, we can ask, “how many pairs of adjacent genes are actually known to be of similar function and is this number statistically significant compared to choosing pairs of genes at random?” Tamames et al. (Tamames, Casari et al. 1997) identified adjacent genes in the complete *H. influenzae* genome and asked if each pair consisted of genes of similar functions. The genes were classified into nine broad functional classes (metabolism, translation, transport, etc.), which allowed the construction of the distribution of all combinations of the nine functional classes. The significance of each combination was estimated using a statistical model comparing the observed associations to those expected by chance alone. Their conclusion was that neighboring genes tend to be functionally linked more often than genes that are not neighbors. Intra-functional doublets (for example, metabolism-metabolism or translation-translation) almost always account for the most significant number of doublets.

Many false positives are detected, however, when querying for functional relationships simply on the basis of proximity on the chromosome. The main problem is that while selective pressures can be inferred to operate on some gene clusters, the constraints on large-scale organization are weak. Thus, while the fact that some genes are neighbors can be quite informative, it can be misleading in other instances. The trick, then, is to isolate those gene pairs whose proximity is significant by filtering out the arbitrary neighbors. Several strategies have been proposed to identify neighboring genes on the chromosome that may be functionally linked, as we discuss in the following sections.

#### Using groups of closely related genomes to identify conserved regions

One method used to ascertain whether the adjacency of a pair of genes in one genome is coincidental or informative makes use of the sequences of other complete genomes to look for conserved gene clusters. In other words, if two genes are consistently adjacent in various genomes, we become confident of an association between them. However, there is an inherent difficulty in choosing the appropriate background. As Dandekar et al. (Dandekar, Snel et al. 1998) point out, if two genomes only diverged recently in evolutionary time, gene clusters in the genomes are expected to be preserved not because they are selectively advantageous but because there has been little time for random drift in the organization of the chromosomes. Conversely, if the reference genome is too distant, then there is less of a chance of finding common orthologs or conserved regions.

The strategy used by Dandekar et al. (Dandekar, Snel et al. 1998) to choose the appropriate phylogenetic background in order to establish the significance of gene proximity is to consider three genomes at a time, and to consider a relation among two proximate genes only if it is conserved among all three. The triplet of genomes is chosen such that the average level of similarity based upon a set of 34 representative orthologs (Huynen and Bork 1998) among two of the three genomes is less than 50%. The group chose three sets of three such genomes. In each triplet of genomes, pairs of genes whose order was conserved were recorded; altogether, among all the sets of genomes, 100 such pairs were detected. This number is considerably smaller than expected and may perhaps be attributed to this strict definition of gene proximity. For example, since *E. coli* has ~4000 protein-coding genes and thus also has ~4000 pairs of adjacent genes, only approximately 2% of the pairs are conserved, assuming that all ~100 pairs of Dandekar et al. were detected in *E. coli*. Thus, these results conclude that a very low degree of gene order is conserved among closely related species.

Most of these gene pairs (75%) correspond to those whose protein products are known to physically interact, for example, ribosomal proteins, RNA polymerase subunits, and transporter subunits. Another 20% correspond to pairs of genes that have been predicted to interact by other methods such as analyses of crystal structures. Another clue towards the notion that the conserved pairs code for functionally related proteins is the observation that sequence similarity is substantially greater among orthologous genes than are members of such pairs than in orthologous genes whose locus along the chromosome is not conserved. If the two proteins interact, then one would expect an

additional selective constraint upon the conservation of complementary structures required for the interaction, thus slowing down the divergence among the orthologs.

### Systematic search for conserved gene order

A more systematic method for identifying consistently paired genes was introduced by Overbeek et al. (Overbeek, Fonstein et al. 1999), who observed that proximate genes are generally functionally related only when they are transcribed in the same direction; indeed, among the 100 pairs of neighboring genes of Dandekar et al., there was only one exception to this rule. Overbeek et. al. therefore defined two genes as proximate if they occur on the same strand and at most 300 bp apart. A pair of proximate genes is only registered if the closest orthologs of the genes in another genome are also proximate. 58,498 pairs are thus linked with a database of 31 genomes, a few of which were not completely sequenced.

The authors take account of similarity of two genomes by using phylogenetic distance according to the standard 16s rRNA tree. For example, if a pair of proximate genes is found in both *H. influenzae* and *E. coli*, it is given a smaller score than a pair found in *E. coli* and *B. subtilis*, because the latter genomes are farther apart than the former. The sum of the scores of all the links that a particular pair of genes X and Y might have made with all the genomes available represents the score for this link.

An important achievement of this analysis is its ability to reconstruct entire pathways, such as purine biosynthesis and glycolysis. The links are clustered such that, within a cluster, each pair of proximate genes is linked to at least one other pair of proximate genes. When a cluster corresponds to the genes of an entire known pathway, this pathway is said to have been reconstructed because the information that the genes form a pathway was not entered but rather was an output of the analysis.

One gene of unknown function, found in many of the genomes considered, was linked to the purine biosynthesis pathway cluster. The hypothesis that this gene participates in purine biosynthesis represents a testable prediction of the functional coupling of the chromosomal proximity method.

One shortcoming of this method is that a proper statistic has not been developed to determine the significance of links among genes based upon chromosomal proximity. The sum rule favors genes that are present in all genomes as opposed to those found only one kingdom, for example.

### Operon detection

An alternative method for identifying genes whose proximity is not arbitrary is the prediction of operons, or multi-gene units of transcription, based only on the genomic sequence, or complemented by other methods such as microarray experiments. As genes within the same operon are likely to be functionally correlated (Pardee 1959), the *in silico* identification of operons offers another method to determine proximate linked genes. By

observing that the genes within an operon are separated by intergenic regions of short, characteristic length, Salgado et al. (Salgado, Moreno-Hagelsieb et al. 2000) were able to identify 75% of the previously known operons, predict many additional operons, and estimate a total of ~700 operons in E.coli. By adding the prerequisite that the genes must be not only neighbors but also members of the same operon, that is, colinear and separated by at most some threshold number of bases, they reduce the risk of linking genes whose proximity is only coincidental.

#### Summary: a close call

Statistically, neighboring genes tend to be functionally related. To prevent reporting false positives, i.e., adjacent genes that are not functionally linked, one should ask whether the genes are also proximate in other genomes. This can be done either systematically by querying all genomes for the gene pair and taking into account the phylogenetic distances among the genomes compared, or by only querying a carefully chosen set of genomes that are at an appropriate distance from the reference genome. Alternately, one could first detect operons and assume that all the comprising genes within an operon are functionally linked. As the sequences of more genomes come to light, the order of genes in a genome will be increasingly telling of associations among neighbors.

## Gene Fusion Analysis

The taxonomy of gene families is complicated by to the existence of “chimeras” (Henikoff, Greene et al. 1997), i.e., genes that are fusions of multiple modules or domains. However, this feature of gene architecture can be extremely useful for functionally linking genes. Thus, for example the distinct and physically distant *E. coli* genes *carB*, *pyrB*, and *pyrC*, whose products act in sequential steps in pyrimidine metabolism, are fused in the human genome.

### Humble beginnings of fusions in bioinformatics

Besides throwing a monkey wrench into the machinery of protein family taxonomies, “chimeras” or fused genes have presented a major challenge for “functionators”, automated functional annotation programs that ascribe putative functions to raw sequences with sequence similarity to known genes. The basic problem arises from a lack of transitive associations among the genes. Consider fusion gene AB, where domains A and B are also found as distinct genes. A is similar in sequence to AB, which in turn is similar to B, but it does not follow that gene A is similar to gene B. Now if A has been studied and determined to be a kinase, uncharacterized gene AB will also be annotated as a kinase. Since B shares a domain with AB, it will likewise be annotated as a kinase. Sequence databases are plagued by such unfortunate situations in which the annotation of a gene is erroneously transferred to a gene that has no sequence similarity.

Ironically, the fusion method suggests that, lack of transitivity notwithstanding, fused pairs, or pairs of genes in one genome that are found fused in another genome, are more likely to be functionally related than unfused pairs. The rationale is that the fusion gene acts as the bridge among the genes, linking them functionally.

### Evidence that distinct genes found fused in another genome are functionally linked

Beginning with *E. coli* genes, Marcotte et al. identified a list of 6809 pairs of non-homologous genes that are found fused in other genomes, and subjected the pairs to three assays for a functional association.

- SwissProt keyword. Through queries for keywords common to the two genes in each of the pairs, 68% of the pairs of known functions have a keyword in common.
- Interacting proteins. A physical interaction between the gene products is a type of functional association. 6.4% of the pairs are composed of genes whose products are known to interact.
- Phylogenetic profiling. 5% of the pairs are also identified by phylogenetic profiling, eight times as many as if the pairs were chosen randomly. That phylogenetic profiling and fusion analysis converge to some degree in their predictions of functional associations increases our confidence that they are indeed identifying real relations.

### Filtering of “promiscuous” domains

Most genes fuse to few partners while a few “promiscuous” genes fuse to many partners. For example, SH3 domains represent over 300 links within 6809 E.coli fusion pairs. To get an idea of the extent of “promiscuity”, consider that when Marcotte et al. removed any link involving a protein with more than 25 links, the number of links drops by 79%.

Is the filtering of these domains from the set of functional links warranted? The rationale for their exclusion is the lack of specificity of information they confer to their link partner. For example, the ATP-binding cassette is represented nearly 200 times in the pairs predicted by Marcotte et al.. When an uncharacterized E. coli gene is found fused to an ATP-binding cassette, the only information gained is that it needs ATP as a source of energy. The degree to which one finds such information useful determines whether filtering promiscuous domains is necessary.

#### Do genes that are found fused elsewhere interact?

To this point, we have discussed linking genes that are found fused elsewhere in the hopes that the link corresponds to a functional association. But exactly what kind of an association is it? Marcotte et al. suggest that the links may hint at physical interactions between the gene products, i.e., the protein products interact with each other as part of their function. Since interactions among proteins are essential to cellular processes, identifying such interactions is an important task, particularly impressive when accomplished at the sequence level.

Marcotte et al. propose a model to explain why a fusion gene is evidence of a physical interaction between the components. Two genes may fuse by a random genetic rearrangement so that they encode a single fusion protein. As the effective concentration of the domains of the protein is now increased by several orders of magnitude, any mutation that leads to or improves a binding surface between the domains reduces the free energy of the protein, and will thus be selected for. When another round of recombination now separates the two domains, they will interact as distinct proteins, since they evolved a binding site while they were fused.

According to this model, the fusion gene is a vestige of the evolution of the interaction between the component genes. However, it is a well-known fact that the average eukaryotic gene is longer than the average prokaryotic gene, suggesting that gene fusion is more common in the gene architecture of eukaryotes. We therefore hypothesize that a fusion may be a beneficial strategy for coping with the cellular complexity of higher organisms by effectively compartmentalizing the component genes. This model predicts a more general functional association between the component genes, rather than, but not precluding, a direct physical interaction.

#### A closer look at the fusion pairs...

Of the 767 fusion pairs found by Marcotte et al. that survived their filtering of “promiscuous” domains, 122 are assigned to metabolic pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata, Goto et al. 1999). Of these, 60

have both members within the same pathway while the remaining 62 do not. Of the former, we identified 34 pairs with distinct functions (i.e., different Enzyme Commission (EC) numbers) that are found fused in a total of 14 genes in other organisms (Yanai, Camacho, and Weng, unpublished results). The remaining 26 pairs have identical EC numbers and mostly correspond to different components of the same complex.

Strikingly, 10 of the 14 fusions are composed of sequential enzymes in their respective metabolic pathway, while another 3 include at most one next-to-nearest neighboring enzymes along the pathway. For example, E.coli genes AroB, AroD, AroE, AroK and AroA, respectively the 2<sup>nd</sup> through 6<sup>th</sup> steps in the shikimate pathway, are all encoded as Aro1 in yeast, though not in this exact order. This suggests that the fusion of genes may be a strategy for ensuring the proximity of the sequential steps of a metabolic pathway. A gene may thus be linked to a certain metabolic pathway, and thus be ascribed a function, if it is found fused to a gene of that pathway. As an example, a regulatory gene may be assigned a putative role in the regulation of the purine biosynthesis if it is also found fused to a gene encoding an enzyme along the purine biosynthesis pathway.

#### Protein-interaction maps by the yeast two-hybrid method

Recent large-scale protein-protein interaction experiments in *S. cerevisiae* (Mewes, Frishman et al. 2000) presented the opportunity to test the ability of the fusion method to detect protein-protein interaction from genomic sequence. 957 protein-protein interactions were identified using the yeast two-hybrid method, almost doubling the known number of physical interactions in *S. cerevisiae*. Using permissive search criteria against a comprehensive sequence database (Holm and Sander 1998), we found that 8 of the 957 (~1%) pairs are composed of proteins that are found fused elsewhere and could have been predicted by the fusion method to interact, 6 of these being self-interactions. By contrast, we detected 1482 pairs of *S. cerevisiae* distinct genes found fused elsewhere but whose products were not found to interact in yeast two-hybrid experiments.

These results indicate that the fusion method has low sensitivity and specificity for detecting direct protein-protein interactions. There are a number of possible explanations for this. Low sensitivity may reflect experimental biases, including the selection of “bait” proteins. Alternatively, the fusion gene for a pair of interacting proteins may not be available, either because it has mutated beyond recognition, or because it has disappeared (Marcotte, Pellegrini et al. 1999), or because it has yet to be sequenced. With respect to the low specificity, we asked whether the fusions primarily indicate functional associations rather than physical interactions.

#### Summary: fusion for function

To systematically test the notion that functional annotations can be transferred transitively, by way of the fusion gene, from one gene to another with no sequence similarity, we examined all pairs of *S. cerevisiae* genes that are found fused together as one gene in another genome (Yanai, Derti and DeLisi, submitted). We associated to each gene the functional category assigned to it by the Munich Information Center for Protein Sequences (MIPS), in which 54% of *S. cerevisiae* genes are mapped to 12 general

categories ((Mewes, Frishman et al. 2000) and note; same as in paper). We then asked for how many of the fusion pairs do both members have functions in common. For 57% of the 1608 fusions of non-homologous genes where both are annotated, we found that the two genes have at least one functional category in common.

The result that genes tend to fuse with genes in the same functional category gives confidence in the concept that fusion analysis will be used to complement sequence analysis. Statistics need to be developed to establish the significance of predictions made by fusion analysis. Thus, for each link made among genes found fused elsewhere, a measure of confidence will also be assigned. This sort of fusion analysis will also give insight into the complex domain architecture that is so ubiquitous, especially in higher organisms.

## Expression profiling, clustering, and regulatory sequences

The leitmotiv of this chapter is the accumulation of clues about the functions of genes through means beyond sequence similarity, and we have thus far considered methods exploiting genomic context to infer functional links among genes. Toward the same end, we now discuss the analysis of large-scale mRNA expression data from microarrays, the search for common upstream regulatory sequences, and the combined use of these methods.

DNA microarrays provide the mRNA levels of thousands of genes simultaneously. With data from many time points and/or conditions, we look for genes with similar expression patterns to infer functional associations. Separately, we can link genes with common upstream regulatory motifs that are putative transcription-factor binding sites. When we combine these methods and find genes with similar expression patterns and common regulatory sequences, we have strong evidence of co-regulation and confidence in the functional links thereby drawn among genes.

### Using DNA microarrays to measure expression levels

DNA microarrays consist of thousands of DNA probe sequences arrayed on a glass or nylon surface, and are used for large-scale gene expression monitoring, *de novo* sequencing or sequence verification. The sample is labelled then hybridized, then the array is scanned. In the case of mRNA, the intensity recorded for each probe is a relative measure of the expression level of a small or large fragment of a gene, with numerous considerations beyond the immediate scope. Whether they are driven by hypothesis or discovery, microarray experiments have produced a torrent of data and led to significant biological findings (Winzeler, Richards et al. 1998; Behr, Wilson et al. 1999; Golub, Slonim et al. 1999; Bittner, Meltzer et al. 2000; Hayward, Derisi et al. 2000; Hughes, Marton et al. 2000; Ly, Lockhart et al. 2000) along with computational and experimental challenges (Aach, Rindone et al. ) and opportunities.

### Limitations of microarrays

Microarrays may not detect transcripts present at low levels, which are likely to include many transcription factors (Thieffry 1999), or faithfully reproduce the dynamic range of the transcripts. Spotted arrays made with long products are particularly susceptible to cross-hybridization, where a target transcript hybridizes to ectopic probes of sufficient similarity. Transcripts dissimilar in sequence can have similar expression patterns (Eisen, Spellman et al. 1998), but this does not preclude cross-hybridization. As we will discuss, expression results and regulatory sequences can potentially provide some clarification in these cases.

Since proteins are the principal agents of the cell, mRNA levels are surrogate indications of gene activity, and there is conflicting evidence regarding the correlation between the two levels (Futcher, Latter et al. 1999; Gygi, Rochon et al. 1999; VanBogelen, Greis et al. 1999). DNA microarrays do not reveal protein interactions or post-transcriptional

modifications, but nevertheless do reveal links not accessible through the study of proteins, in part because transcription and translation are uncoupled in eukaryotes.

### Clustering microarray expression data

A common analysis of microarray expression data consists of clustering expression profiles, whether from a time-course experiment or multiple experiments, in order to find genes with similar profiles. A multitude of approaches have been developed or adapted for this purpose (Sherlock 2000). The profiles are typically normalized first, so that clusters contain patterns whose absolute levels may differ but which have similar shapes. The general procedure is to select genes with significant changes in expression, then to establish a measure of similarity between expression patterns, such as correlation (Eisen, Spellman et al. 1998), Euclidean distance (Tavazoie, Hughes et al. 1999) or mutual information (Herwig, Poustka et al. 1999), then to group similar patterns. The latter implies the subjective and iterative selection of a threshold for similarity.

In hierarchical clustering (Eisen, Spellman et al. 1998), the two most similar patterns are clustered and replaced by an average, and the other patterns are added successively by decreasing order of similarity. The result resembles a phylogenetic tree, and clusters are obtained by setting a threshold at some level in the hierarchy. In self-organizing maps (SOMs) (Tamayo, Slonim et al. 1999), a number of reference nodes are dispersed initially in expression space in a geometric pattern such as a grid. At each iteration, a gene is chosen at random and each node is moved toward that gene's expression vector by a distance proportional to its closeness with that vector, but this distance decreases as the iterations progress. After many iterations, each expression pattern is assigned to the closest node.

Through an iterative reallocation of the members, k-means clustering (Tavazoie, Hughes et al. 1999) minimizes intra-cluster dispersion for a specified number (k) of clusters. The reference vectors are initialized randomly, as with SOMs, but each gene is assigned to its closest reference vector, and only that vector is readjusted to represent an average (means) of the expression profiles in its cluster. Iterations are halted when no expression patterns are reallocated among clusters. Herwig et al. modified the method so that close clusters can merge. Unlike SOMs and k-means clustering, hierarchical clustering is deterministic (Sherlock 2000).

Heyer et al. (Heyer, Kruglyak et al. 1999) developed a clustering algorithm specifically for expression patterns. A single pair of corresponding data points can yield a strong correlation between two otherwise dissimilar profiles if those points are strong outliers, and the authors propose the use of the jackknife correlation to avoid this problem. This value is obtained as follows for two profiles consisting of N data points: for  $i = 1$  to N, omit the  $i$ th data point from each profile, calculate the correlation between the profiles, then restore the  $i$ th data points and increment  $i$ ; the worst possible correlation among these is the jackknife correlation. If two points were omitted from each profile at each iteration, the result would be the second-order jackknife correlation. Unlike other methods where the order of clustering can affect the results, the algorithm of Heyer et al.

exhaustively uses each gene as the seed for a cluster, searching for the greatest number of genes within a similarity threshold, then repeats the process for the remaining genes. Of particular interest to this chapter is the comment by these authors regarding transitivity: if we are to infer a functional relation between genes in a cluster, then the genes within a cluster must be strongly similar to each other, and not related through multiple links.

This exhaustive search for the worst measure of similarity between expression patterns can be extended from correlation to other measures of similarity or distance. In addition, if the same time point consistently leads to the worst correlation, perhaps it is noisy and should be excluded. Lastly, it is possible that some genes exhibit similar expression patterns in some conditions but not others, and this jack-knife process also leads to the idea of systematically or randomly omitting data points from various conditions to find the subsets where genes exhibit similar expression patterns.

### Searching for proximal regulatory sequences

An important point of control of gene expression is the regulation of transcription initiation. In simple terms, *trans*-acting transcription factors bind to *cis*-acting regulatory sequences upstream of genes and either enhance or repress transcription, directly or through accessory factors. The search for regulatory sites in genomic sequences is consequently subject to much attention (Fickett and Wasserman 2000; Stormo 2000).

Pursuing the idea of cooperative binding, Wagner (Wagner 1999) searched for conserved clusters of multiple sites in *S. cerevisiae*. Based on the presumed conservation of functional sequences despite variations in neighboring non-coding fragments, Miller and colleagues have reported the utility of looking for sequences conserved among the promoter regions of human and rodent homologs (Hardison, Oeltjen et al. 1997; Oeltjen, Malley et al. 1997; Ansari-Lari, Oeltjen et al. 1998; Mallon, Platzer et al. 2000) toward the identification of regulatory sites. A number of methods search for a common site in multiple unaligned sequences, typically upstream regions of genes thought to be co-regulated (Roth, Hughes et al. 1998; van Helden, Andre et al. 1998; Stormo 2000).

Regulatory sequences can be represented, among others, as consensus sequences or weight matrices, the latter being able to detect more sites and to extend binding sequences beyond the consensus (P. Estep and J. Hughes, personal communication). In a weight matrix, each position is indicative of the occurrence of each base and, optionally, of the information provided, which depends upon the GC content of the sequences being considered. Roulet et al. (Roulet, Bucher et al. 2000) note that neither weight matrices nor consensus sequences adequately represent the binding of the CTF/NFI family of eukaryotic transcription factors, which bind to two half sites separated by a variable region, but can affect transcription through a single half site. The authors introduce a profile that accommodates these aspects but still treats each position as independent.

Stormo (Stormo 2000) notes that a limitation of weight matrices is that each position is assumed to contribute additively, i.e., independently. Brunak and colleagues (Baldi, Chauvin et al. 1998; Pedersen, Baldi et al. 1999) have shown that the use of structural

information beyond sequence, such as dinucleotide stacking energies and helix bendability, leads to profiles with recurrent changes from less bendable upstream of transcription start to more bendable downstream. It is possible that this guides the formation of a nucleosome downstream of transcription start, leaving the region just upstream accessible to transcription factors. This can also help delineate genes (Baldi, Chauvin et al. 1998), and in another example of combined approaches, the comparison of paralogs and the consideration of structural profiles can corroborate each other in the simultaneous search for genes and upstream regulatory sequences.

### Combining microarray expression data with regulatory sequence analysis

As Heyer et al. note (Heyer, Kruglyak et al. 1999), a relationship is expected between coexpression and coregulation, and a specific example in *S. cerevisiae* is illustrative. With data from multiple microarray experiments, Eisen et al. (Eisen, Spellman et al. 1998) found one cluster containing many genes for structural and regulatory subunits of the proteasome, a large and abundant complex that degrades cytosolic proteins once they are ubiquitinated. Mannhaupt et al. (Mannhaupt, Schnell et al. 1999) observed the motif 5'-GGTGGCAA-3' upstream of a number of genes, including 26 proteasome genes. They isolated the protein that binds to this motif, Rpn4p, and confirmed its ability to promote the transcription of reporter constructs. In a search for regulatory motifs in *S. cerevisiae* genes grouped only by functional annotation, Hughes et al. (Hughes, Estep et al. 2000) found, among others, the same motif upstream of 31 proteasome genes and a number of related genes. They independently isolated Rpn4p by one-hybrid, and confirmed its activity by mRNA expression after knocking out and overexpressing Rpn4p.

### Analysis of *S. cerevisiae* data sets, a short history

The analysis of genome-wide expression in yeast has become increasingly sophisticated, and we briefly follow this evolution.

Roth et al. (Roth, Hughes et al. 1998) measured expression of all *S. cerevisiae* genes in heat shock, mating type and galactose response, taking single time points and looking for regulatory motifs in a number of the most up- and down-regulated genes. Cho et al. (Cho, Campbell et al. 1998) and Spellman et al. (Spellman, Sherlock et al. 1998) used synchronized cells and assayed expression throughout two cell cycles. Cho et al. clustered the genes visually by the stage of the cell cycle in which they peaked (early G1, late G1, S, G2, M), identified 420 as exhibiting cell-cycle-dependent periodicity, and searched for hexa- and heptanucleotides overrepresented in the 500 bases upstream of translation start sites in every gene by cluster. They noticed moderate though statistically significant co-expression of directly adjacent genes, and found evidence of coordinate degradation of mRNA and proteins. Continuing this study in detail, Wolfsberg et al. (Wolfsberg, Gabrielian et al. 1999) found 9 hexamers and 12 pentamers overrepresented in the upstream sequences of these 420 transcripts, examining preferences in orientation and distance from the translation start site. They searched for motifs with and without position dependence, and found some new motifs and some known motifs, though

admittedly not whole binding sites. If an element is overrepresented but its reverse complement is not, the authors hypothesize that the site is orientation-dependent.

Spellman et al. grouped genes by phase of peak expression and by similarity of expression patterns using the approach of Eisen et al, and used Fourier analysis to look for periodicity. They identified 800 genes as being differentially regulated in the cell cycle, 304 of them in common with Cho et al., attributing the differences in part to experimental artifacts. They then looked for consensus motifs, optionally with specific mismatches, in the 700 bases upstream of genes. Their strongest cluster included 9 histone genes known to be regulated in part by mRNA degradation. Incidentally, histone genes were the first in *S. cerevisiae* for which periodic regulation in the cell division cycle was observed (Hereford, Osley et al. 1981). The authors were unable to find motifs in some of the looser clusters, and observe that the control of stability may explain some fluctuations in mRNA levels. They did find a motif associated with both response to glucose and the cell cycle, perhaps linked by the switch from stationary phase to growth.

With the intent of developing a systematic and more broadly applicable approach not biased by knowledge of the biology, Tavazoie et al. examined the data of Cho et al. and clustered the genes only by their expression profiles, that is, ignoring periodicity and landmarks. They clustered the 3000 most variable ORFs, then used AlignACE (Roth, Hughes et al. 1998) to search for upstream regulatory sequences in each of the resulting 30 clusters. This software uses Gibbs sampling to simultaneously align sequences and find regulatory subsequences, and constructs matrices to describe motifs, rather than consensus sequences. The authors found 18 significant motifs for 12 clusters, seven of which had been found experimentally and were known to regulate a number of genes in their respective clusters. They too note the control of mRNA stability as one explanation for clusters without strong motifs. By searching the genome, they found the motifs to be highly specific for their clusters. Interestingly, the previously known motifs for groups of known co-regulated genes were consistently ranked highest by the algorithm, conferring credibility to the new motifs. In addition, there was a strong correlation between the tightness of a cluster and the presence of a significant motif, and genes with significant motifs tended to be closer to the center of their clusters. In related papers, Rouillon et al. and Patton et al. (Patton, Peyraud et al. 2000; Rouillon, Barbey et al. 2000) refer to the findings of Spellman et al. and Tavazoie et al. about the involvement of MET genes in the cell cycle as clues toward the link between sulfur amino acid metabolism, ubiquitin degradation and the control of the cell cycle.

In an article that bridges various sections of this chapter and illustrates the utility of combining multiple methods, Cohen et al. (Cohen, Mitra et al. 2000) examined the data of Cho et al. and found a significant correlation between adjacency on the chromosome and correlated expression levels. This effect was independent of orientation and statistically significant compared to non-adjacent genes. Although many adjacent pairs did not exhibit correlated expression, the effect did apply to pairs and triplets of adjacent genes throughout the genome, with fewer than 5% of the most correlated adjacent pairs consisting of homologs, and with an apparent decline with increasing intergenic distance.

### Conclusion of expression

Niehrs and Pollet (Niehrs and Pollet 1999) remark that groups of genes with coordinate expression levels in eukaryotes are analogous to operons in prokaryotes, while Heyer et al. note that a relationship is expected between coexpression and coregulation. Groups of coregulated eukaryotic genes are more difficult to identify than operons, but similar expression levels and common regulatory sequences corroborate each other in establishing coregulation. When they disagree, meaning that co-expressed genes do not have common regulatory sequences or genes with common regulatory sequences are not co-expressed, we can speculate about the possible reasons: cross-hybridization (genes only appear to be co-expressed due to experimental artifacts), incidental expression of neighboring genes, mRNA degradation (genes are activated together, but differential degradation rates obfuscate the expression profiles), silencing of entire chromosomal regions in higher organisms and other cross-reactions. It is also possible that co-expressed genes without a common regulatory element are in fact indirectly co-regulated in that they are subject to a “master controller” which governs more than one regulatory element.

Most of the publicly available expression data has been generated for *S. cerevisiae*, and the development of these computational methods in this well-characterized organism will be of immense value when extended to expression studies in the many sequenced organisms about which much less is known. In addition, with the advent of large-scale expression data for *E. coli* (Richmond, Glasner et al. 1999), in which transcription and translation are coupled, it is likely that greater advances will be made in determining causality in time-course expression patterns, facilitating the biological elucidation and computational inferences of genetic networks.

## Integration

As the methods described in this chapter, essentially serve to link one gene to another, and therefore the output of all the method are comparable to one another. Furthermore, as these methods all link proteins based upon different principles they can strengthen and complement each other. Whereas each method is expected to derive only a small subset of all possible functional links among the genes of a genome, the links may be compiled together, to both strengthen each link, if it is identified by many different methods or complement each other by connecting together as many genes as possible. Thus in attempting to identify all functional links among a genome's genes, it is best to use all possible methods to link genes.

Marcotte et al. combined three of the methods described here, phylogenetic profiling, domain fusion analysis, and correlated mRNA expression, in attempt to link together the 6,217 genes of the *S. cerevisiae* genome. Using the keyword annotations that are available at SwissProt database, the group could assess the accuracy of prediction of each of the three methods. In other words, if two proteins with available keyword annotations in the SwissProt database are linked by phylogenetic profiling for example, then one can verify if the two proteins do indeed share a keyword in common and thus a functional association. From this analysis different weights could be assigned to each method to reflect the confidence one has in their predictions. Links identified by two or three links were given highest confidence, links made by phylogenetic profiling were assigned high confidence and all the rest were these links made by the domain fusion method or correlated mRNA expression. Of the 2,557 genes in *S. cerevisiae* that have unassigned function, 374 could be assigned a general function based upon the high and highest confidence links. If all links are considered, then the group can assign a function to 1,589 genes.

Methods to predict a gene's function from a consideration of its links to other genes represent an exciting component of the emerging field of functional genomics. Here we reviewed methods distinct from classical sequence similarity: link by common evolutionary patterns, by consistent proximity on the chromosomes of different genomes, different domain architecture, and correlated expression. We believe that codon usage will also be a useful method for functional linkage (Medigue, Rouxel et al. 1991) as well as an *in libro* (Andrade and Bork 2000) consideration of co-occurrence of genes in publications, or "publons"- units of literature. As these methods accumulate, the computational end of functional genomics is becoming an increasingly crucial tool in the transformation of the genetic information to biological knowledge.

## Sequence Analysis *Without* Sequence Similarity

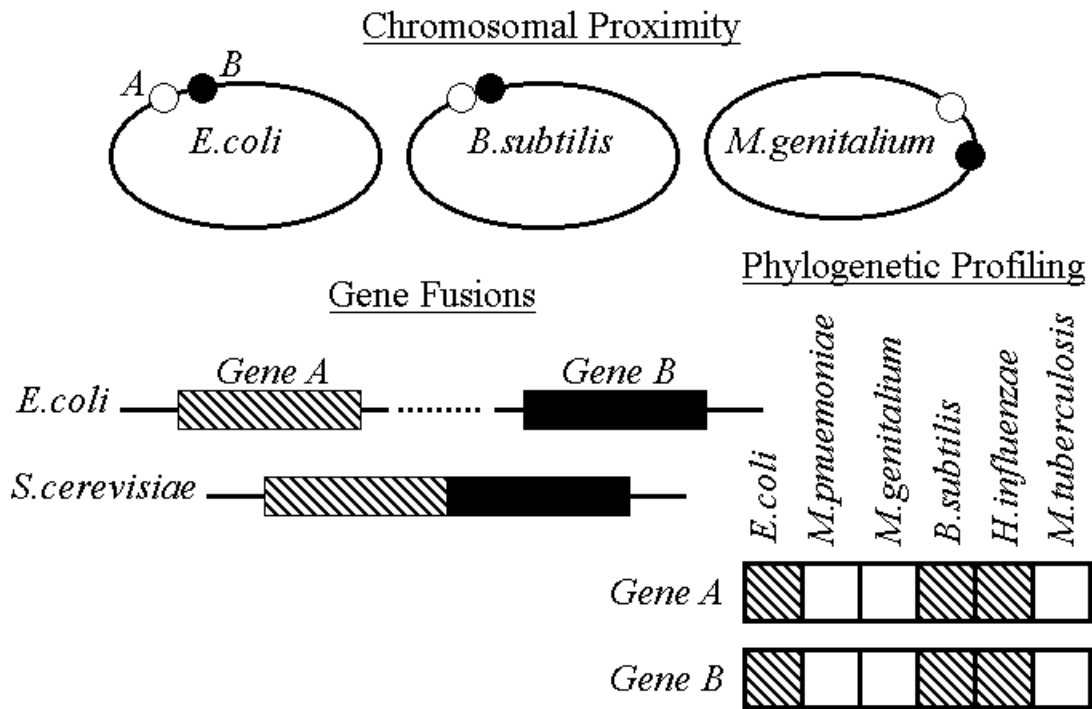


Figure 1: Non-homology based functional prediction methods. In chromosomal proximity genes that systematically neighbors across genomes are linked. Two genes may also be linked, using fusion analysis, if they found fused as one multidomain gene in another genome. The phylogenetic profiling method links genes that are either both present or both absent across genomes.

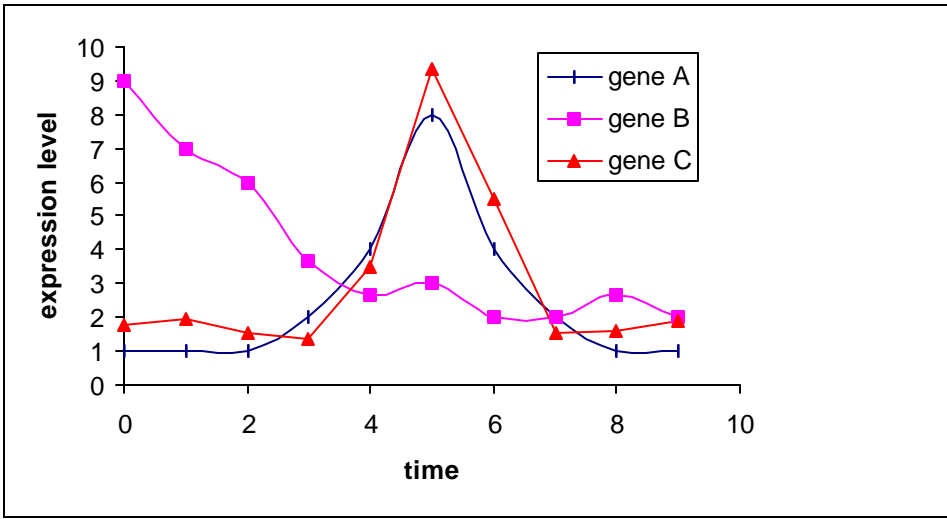


Figure 2: Linking genes by expression profiling. Genes with similar expression profiles (genes A and C) are potentially co-regulated. If they also have a common upstream regulatory region, we have greater confidence in this hypothesis.

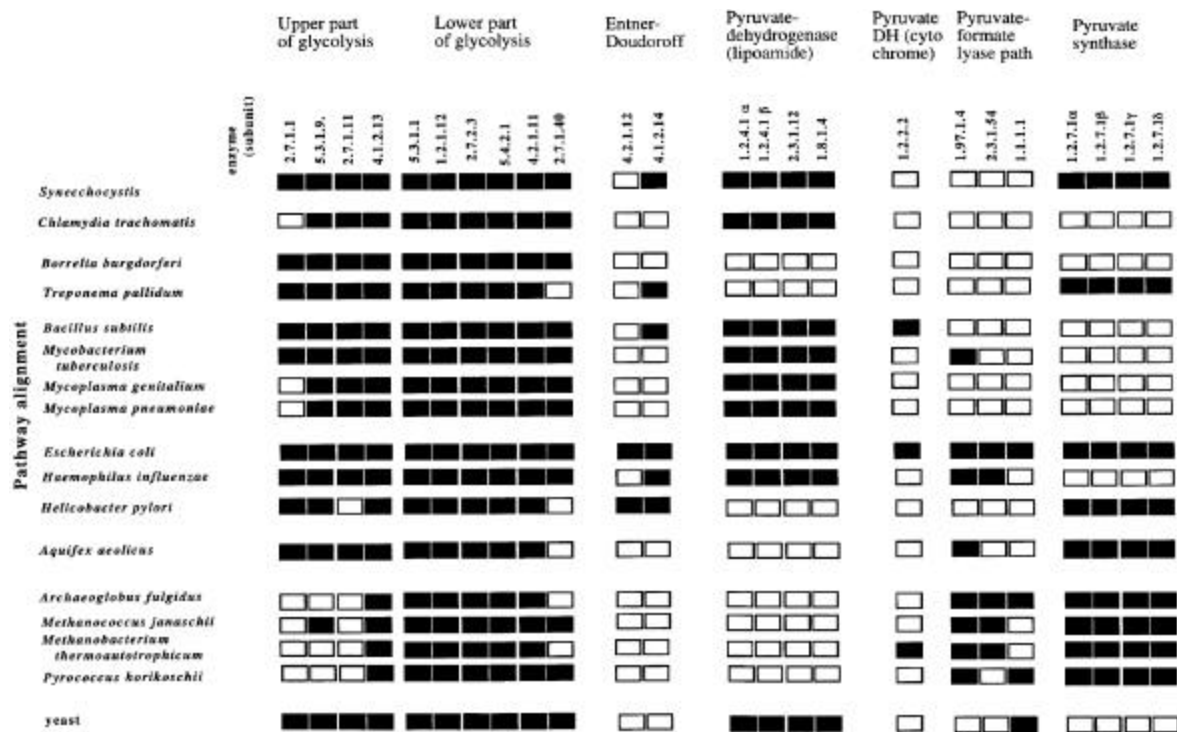


Figure 2: Pathway alignment for glycolysis, Entner-Doudoroff pathway and pyruvate processing. Enzymes for each pathway part (top; EC numbers and enzyme subunits are given below these) are compared in 17 organisms and represented as small rectangles. Filled and empty rectangles indicate the presence and absence respectively of enzyme-encoding genes in the different species listed at the left. Kindly provided by Thomas Dandekar (Dandekar, Schuster et al. 1999).

## References:

- Aach, J., W. Rindone, et al. "Systematic management and analysis of yeast gene expression data." .
- Ansari-Lari, M. A., J. C. Oeltjen, et al. (1998). "Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6." Genome Res **8**(1): 29-40.
- Baldi, P., Y. Chauvin, et al. (1998). "Computational applications of DNA structural scales." Ismb **6**: 35-42.
- Behr, M. A., M. A. Wilson, et al. (1999). "Comparative genomics of BCG vaccines by whole-genome DNA microarray." Science **284**(5419): 1520-3.
- Bittner, M., P. Meltzer, et al. (2000). "Molecular classification of cutaneous malignant melanoma by gene expression profiling." Nature **406**(6795): 536-40.
- Chiu IM, R. E., Givol D, Robbins KC, Tronick SR, Aaronson SA. (1984). "Nucleotide sequence analysis identifies the human c-sis proto-oncogene (simian sarcoma virus transforming gene) as a structural gene for platelet-derived growth factor." Cell **37**(1): 123-129.
- Cho, R. J., M. J. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell **2**(1): 65-73.
- Cohen, B. A., R. D. Mitra, et al. (2000). "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression." Submitted.
- Danchin, A. (1998). "The Delphic boat or what the genomic texts tell us." Bioinformatics **14**(5): 383.
- Dandekar, T., S. Schuster, et al. (1999). "Pathway alignment: application to the comparative analysis of glycolytic enzymes." Biochem J **343 Pt 1**: 115-24.
- Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem Sci **23**(9): 324-8.
- Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.
- Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-6.
- Fickett, J. W. and W. W. Wasserman (2000). "Discovery and modeling of transcriptional regulatory regions." Curr Opin Biotechnol **11**(1): 19-24.
- Galperin, M. Y. and E. V. Koonin (2000). "Who's your neighbor? new computational approaches for functional genomics" Nat Biotechnol **18**(6): 609-13.
- Golub, T. R., D. K. Slonim, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science **286**(5439): 531-7.
- Gygi, S. P., Y. Rochon, et al. (1999). "Correlation between protein and mRNA abundance in yeast." Mol Cell Biol **19**(3): 1720-30.
- Hardison, R. C., J. Oeltjen, et al. (1997). "Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome." Genome Res **7**(10): 959-66.
- Hayward, R. E., J. L. Derisi, et al. (2000). "Shotgun DNA microarrays and stage-specific gene expression in Plasmodium falciparum malaria." Mol Microbiol **35**(1): 6-14.

Henikoff, S., E. A. Greene, et al. (1997). "Gene families: the taxonomy of protein paralogs and chimeras." Science **278**(5338): 609-14.

Hereford, L. M., M. A. Osley, et al. (1981). "Cell-cycle regulation of yeast histone mRNA." Cell **24**(2): 367-75.

Herwig, R., A. J. Poustka, et al. (1999). "Large-scale clustering of cDNA-fingerprinting data." Genome Res **9**(11): 1093-105.

Heyer, L. J., S. Kruglyak, et al. (1999). "Exploring expression data: identification and analysis of coexpressed genes." Genome Res **9**(11): 1106-15.

Holm, L. and C. Sander (1998). "Removing near-neighbour redundancy from large protein sequence collections." Bioinformatics **14**(5): 423-9.

Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*." J Mol Biol **296**(5): 1205-14.

Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-26.

Huynen, M., T. Dandekar, et al. (1998). "Differential genome analysis applied to the species-specific features of *Helicobacter pylori*." FEBS Lett **426**(1): 1-5.

Huynen, M., B. Snel, et al. (2000). "Exploitation of gene context" Curr Opin Struct Biol **10**(3): 366-70.

Huynen, M. A. and P. Bork (1998). "Measuring genome evolution." Proc Natl Acad Sci U S A **95**(11): 5849-56.

Huynen, M. A., T. Dandekar, et al. (1999). "Variation and evolution of the citric-acid cycle: a genomic perspective." Trends Microbiol **7**(7): 281-91.

Karzai, A. W., M. M. Susskind, et al. (1999). "SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA)." Embo J **18**(13): 3793-9.

Koonin, E. V., A. R. Mushegian, et al. (1997). "Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea." Mol Microbiol **25**(4): 619-37.

Lawrence, J. G. (1997). "Selfish operons and speciation by gene transfer." Trends Microbiol **5**(9): 355-9.

Ly, D. H., D. J. Lockhart, et al. (2000). "Mitotic misregulation and human aging." Science **287**(5462): 2486-92.

Mallon, A. M., M. Platzer, et al. (2000). "Comparative genome sequence analysis of the Bpa/Str region in mouse and Man." Genome Res **10**(6): 758-75.

Mannhaupt, G., R. Schnall, et al. (1999). "Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast." FEBS Lett **450**(1-2): 27-34.

Marcotte, E. M. (2000). "Computational genetics: finding protein function by nonhomology methods" Curr Opin Struct Biol **10**(3): 359-65.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-3.

Mewes, H. W., D. Frishman, et al. (2000). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **28**(1): 37-40.

Mushegian, A. R. and E. V. Koonin (1996). "A minimal gene set for cellular life derived by comparison of complete bacterial genomes" Proc Natl Acad Sci U S A **93**(19): 10268-73.

Niehrs, C. and N. Pollet (1999). "Synexpression groups in eukaryotes." Nature **402**(6761): 483-7.

Oeltjen, J. C., T. M. Malley, et al. (1997). "Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains." Genome Res **7**(4): 315-29.

Ogata, H., S. Goto, et al. (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Res **27**(1): 29-34.

Ouzounis, C. and N. Kyrpides (1996). "The emergence of major cellular processes in evolution." FEBS Lett **390**(2): 119-23.

Overbeek, R., M. Fonstein, et al. (1999). "The use of gene clusters to infer functional coupling." Proc Natl Acad Sci U S A **96**(6): 2896-901.

Pardee, A. B., F. Jacob, J. Monod (1959). "The genetic control and cytoplasmic expression of "inducibility" in the synthesis of B-galactosidase by E. coli." J.Mol.Biol. **1**: 165-178.

Patton, E. E., C. Peyraud, et al. (2000). "SCF(Met30)-mediated control of the transcriptional activator Met4 is required for the G(1)-S transition." Embo J **19**(7): 1613-24.

Pedersen, A. G., P. Baldi, et al. (1999). "The biology of eukaryotic promoter prediction- a review." Comput Chem **23**(3-4): 191-207.

Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.

Richmond, C. S., J. D. Glasner, et al. (1999). "Genome-wide expression profiling in Escherichia coli K-12." Nucleic Acids Res **27**(19): 3821-35.

Roth, F. P., J. D. Hughes, et al. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol **16**(10): 939-45.

Rouillon, A., R. Barbey, et al. (2000). "Feedback-regulated degradation of the transcriptional activator Met4 is triggered by the SCF(Met30 )complex." Embo J **19**(2): 282-94.

Roulet, E., P. Bucher, et al. (2000). "Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites." J Mol Biol **297**(4): 833-48.

Salgado, H., G. Moreno-Hagelsieb, et al. (2000). "Operons in Escherichia coli: genomic analyses and predictions." Proc Natl Acad Sci U S A **97**(12): 6652-7.

Sherlock, G. (2000). "Analysis of large-scale gene expression data." Curr Opin Immunol **12**(2): 201-5.

Snel, B., P. Bork, et al. (1999). "Genome phylogeny based on gene content." Nat Genet **21**(1): 108-10.

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Tamames, J., G. Casari, et al. (1997). "Conserved clusters of functionally related genes in two bacterial genomes." J Mol Evol **44**(1): 66-73.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-12.

Tatusov, R. L., M. Y. Galperin, et al. (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Res **28**(1): 33-6.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-7.

Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture" Nat Genet **22**(3): 281-5.

Tekaia, F., A. Lazcano, et al. (1999). "The genomic tree as revealed from whole proteome comparisons." Genome Res **9**(6): 550-7.

Thieffry, D. (1999). "From global expression data to gene networks." Bioessays **21**(11): 895-9.

van Helden, J., B. Andre, et al. (1998). "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." J Mol Biol **281**(5): 827-42.

VanBogelen, R. A., K. D. Greis, et al. (1999). "Mapping regulatory networks in microbial cells." Trends Microbiol **7**(8): 320-8.

Wagner, A. (1999). "Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes." Bioinformatics **15**(10): 776-84.

Watanabe, H., H. Mori, et al. (1997). "Genome plasticity as a paradigm of eubacteria evolution." J Mol Evol **44**(Suppl 1): S57-64.

Weinstein, J. N. (1998). "Fishing expeditions." Science **282**(5389): 628-9.

Winzler, E. A., D. R. Richards, et al. (1998). "Direct allelic variation scanning of the yeast genome." Science **281**(5380): 1194-7.

Wolfsberg, T. G., A. E. Gabrielian, et al. (1999). "Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*." Genome Res **9**(8): 775-92.