15

COMPARATIVE GENOME ANALYSIS

Michael Y. Galperin and Eugene V. Koonin

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland

The first complete genome sequences of cellular life forms have become available in just the last several years. In 1995, the genomes of the first two bacteria, *Haemophilus influenzae* and *Mycoplasma genitalium*, were reported (Fleischmann et al., 1995; Fraser et al., 1995). One year later, the first archaeal (*Methanococcus jannaschii*) and the first eukaryotic (yeast *Saccharomyces cerevisiae*) genomes were completed (Bult et al., 1996; Goffeau et al., 1996). 1997 was marked by a landmark achievement—the sequencing of the genomes of the two best-studied bacteria, *Escherichia coli* (Blattner et al., 1997) and *Bacillus subtilis* (Kunst et al., 1997). Many more bacterial and archaeal genomes, as well as the first genome of a multicellular eukaryote, the nematode *Caenorhabiditis elegans*, have since been sequenced (see details below), providing ample material for comparative analysis.

A notable (and perhaps disappointing to many biologists) outcome of these first genome projects is that at least one-third of the genes encoded in each genome had no known or predictable function; for many of the remaining genes, only a general functional prediction appeared possible. The depth of our ignorance becomes particularly obvious on examination of the genome of *Escherichia coli* K12, arguably the most extensively studied organism among both prokaryotes and eukaryotes. Even in this all-time favorite model organism of molecular biologists, at least 40% of the genes have unknown function (Koonin, 1997). On the other hand, it turned out that the level of evolutionary conservation of microbial proteins is rather uniform, with ~70% of gene products from each of the sequenced genomes having homologs in distant genomes (Koonin et al., 1997). Thus, the functions of many of these genes

can be predicted simply by comparing different genomes and by transferring functional annotation of proteins from better-studied organisms to their orthologs from lesser-studied organisms. This makes comparative genomics a powerful approach for achieving a better understanding of the genomes and, subsequently, of the biology of the respective organisms. Here, we describe databases that store genomic information and bioinformatics tools that are used in the computational analysis of complete genomes. The subject of comparative genomics includes a number of distinct aspects, and it is unrealistic to cover them all in a brief chapter. We limit the discussion to the analysis of protein sets from completely sequenced genomes. Because most of the latter are from prokaryotes, there is an inevitable focus on prokaryotic biology in the presentation. Furthermore, in our choice of the genome analysis tools to discuss in detail, we decided to concentrate largely on Web-based ones that are readily accessible to any user, as opposed to stand-alone software that has more limited applicability.

PROGRESS IN GENOME SEQUENCING

By the beginning of 2000, genomes of 23 different unicellular organisms (5 archaeal, 17 bacterial, and 1 eukaryotic) had been completely sequenced. At least 70 more microbial genomes were in different stages of completion with respect to sequencing. Periodically updated lists of both finished and unfinished publicly funded genome sequencing projects are available in the GenBank Entrez Genomes division and at the site maintained by The Institute for Genome Research (TIGR) and at Integrated Genomics. A complete list of sequencing centers world-wide can be found at the NHGRI Web site. One can retrieve the actual sequence data from the NCBI FTP site or from the FTP sites of each individual sequencing center. A convenient sequence retrieval system is maintained also at the DNA Data Bank of Japan. In the framework of the Reference Sequences (RefSeq) project, NCBI has recently started to supplement the lists of gene products with some valuable sequence analysis information, such as the lists of best hits in different taxa, predicted functions for uncharacterized gene products, frame-shifted proteins, and the like. On the other hand, sequencing centers like TIGR have been updating their sequence data, correcting some of the sequencingerrors and, accordingly, their sites may contain more recent data on unfinished genome sequences.

General-Purpose Databases for Comparative Genomics

Because the World Wide Web makes genome sequences available to anyone with Internet access, there exists a variety of databases that offer more or less convenient access to basically the same sequence data. However, several research groups, specializing in genome analysis, maintain databases that provide important additional information, such as operon organization, functional predictions, three-dimensional structure, and metabolic reconstructions.

PEDANT. This useful Web resource provides answers to most standard questions in genome comparison (Frishman and Mewes, 1997). PEDANT provides an easy way to ask simple questions, such as finding out how many proteins in *H. pylori* have known (or confidently predicted) three-dimensional structures or how many

NAD^+-dependent alcohol dehydrogenases (EC 1.1.1.1) are encoded in the *C. elegans* genome (Fig. 15.1). The list of standard PEDANT queries includes EC numbers, PROSITE patterns, Pfam domains, BLOCKS, and SCOP domains, as well as PIR keywords and PIR superfamilies. Although PEDANT does not allow the users to enter their own queries, the variety of data available at this Web site makes it a convenient entry point into the field of comparative genome analysis.

COGs. The Clusters of Orthologous Groups (COGs) database has been designed to simplify evolutionary studies of complete genomes and improve functional assignments of individual proteins (Tatusov et al., 1997, 2000). It consists of ~2,800 conserved families of proteins (COGs) from each of the completely sequenced genomes. Each COG contains orthologous sets of proteins from at least three phylogenetic lineages, which are assumed to have evolved from an individual ancestral protein. By definition, *orthologs* are genes that are connected by vertical evolutionary descent (the ''same'' gene in different species) as opposed to *paralogs*—genes related by duplication *within* a genome (Fitch, 1970; Henikoff et al., 1997). Because orthologs typically perform the same function in all organisms, delineation of orthologous families from diverse species allows the transfer of functional annotation from better-studied organisms to the lesser-studied ones. The protein families in the COG database are separated into 17 functional groups that include a group of uncharacterized yet conserved proteins, as well as a group of proteins for which only a general functional assignment appeared appropriate (Fig. 15.2). This site is particularly useful for functional predictions in borderline cases, where protein similarity levels are fairly low. Due to the diversity of proteins in COGs, sequence similarity searches against the COG database can often suggest a possible function for a protein that otherwise has no clear database hits. This database also offers some convenient tools for a comparative analysis of complete genomes as will be described below.

KEGG. The Kyoto Encyclopedia of Genes and Genomes (KEGG) centers around cellular metabolism (Kanehisa and Goto, 2000). This Web site presents a comprehensive set of metabolic pathway charts, both general and specific, for each of the completely-sequenced genomes, as well as for *Schizosaccharomyces pombe, Arabidopsis thaliana, Drosophila melanogaster*, mouse, and human. Enzymes that are already identified in a particular organism are color-coded, so that one can easily trace the pathways that are likely to be present or absent in a given organism (Fig. 15.3). For the metabolic pathways covered in KEGG, lists of orthologous genes that code for the enzymes participating in these pathways are also provided. It is also indicated whenever these genes are adjacent, forming likely operons. A very convenient search tool allows the user to compare two complete genomes and identify all cases in which conserved genes in both organisms are adjacent or located relatively close (within 5 genes) to each other. The KEGG site is continuously updated and serves as an ultimate source of data for the analysis of metabolism in various organisms.

MBGD. The Microbial Genome Database (MBGD) at the University of Tokyo offers another convenient tool for searching for likely homologs among all sequenced microbial genomes. In contrast to COGs, MBGD assigns homology relationships based solely on sequence similarity (BLASTP values of 10^{-2} or less). MBGD allows the user to submit several sequences at once (up to 2,000 residues) for searching

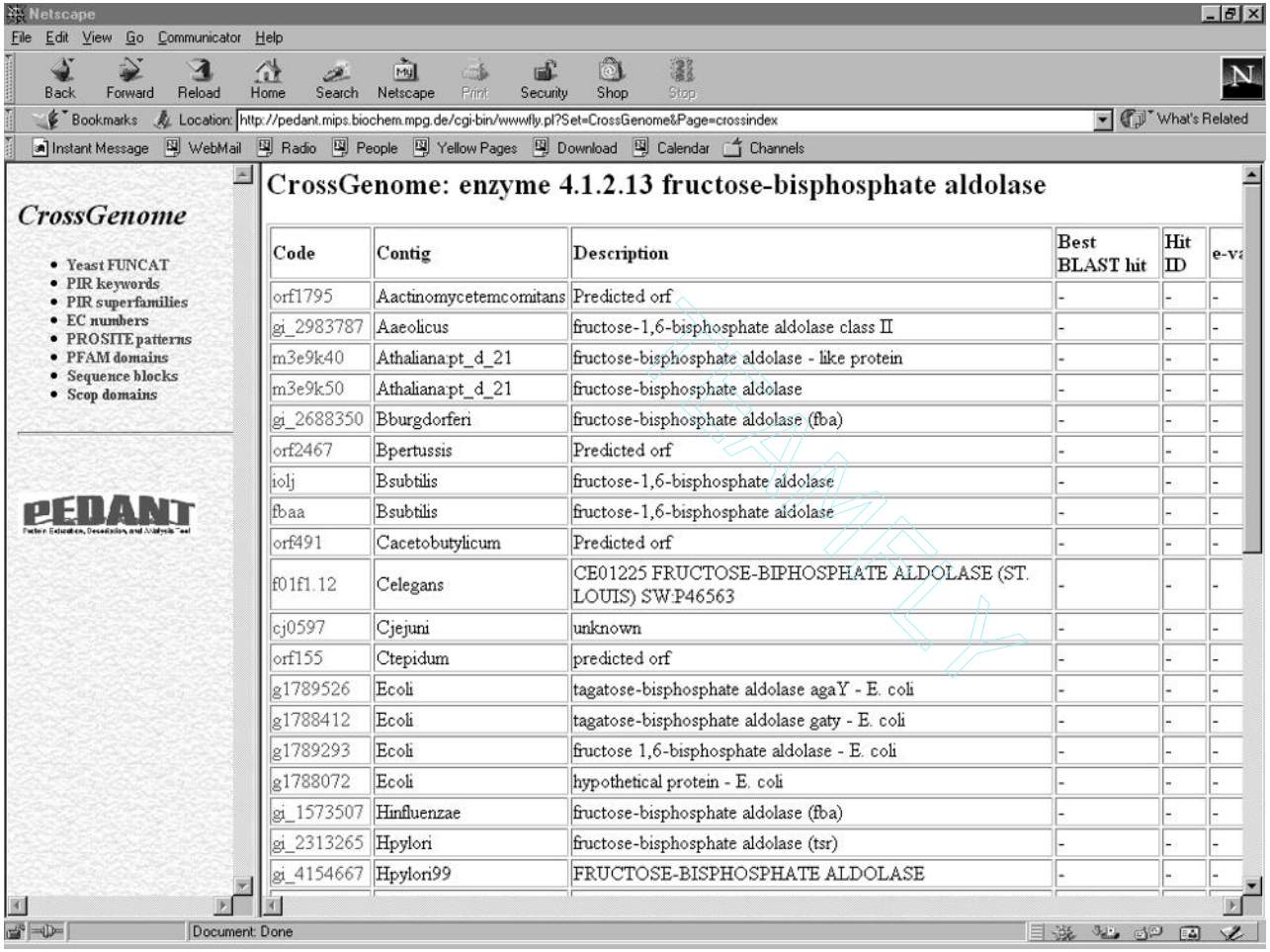


Figure 15.1. The list of fructose-1,6-bisphosphate aldolases in different genomes, as generated by PEDANT. The left frame lists standard PEDANT queries.

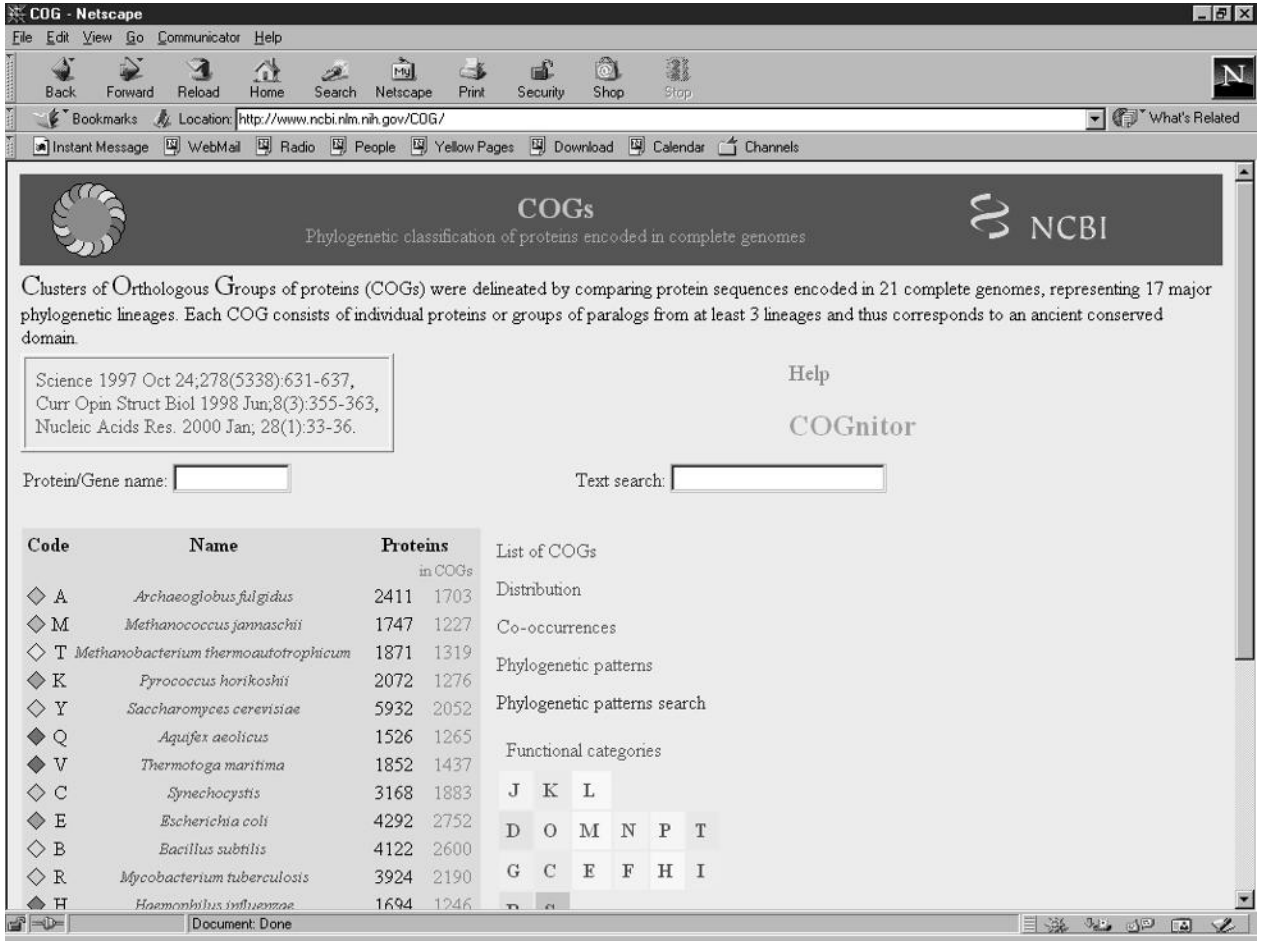


Figure 15.2. The home page of the Clusters of Orthologous Groups of proteins (COGs) database. See text for details.

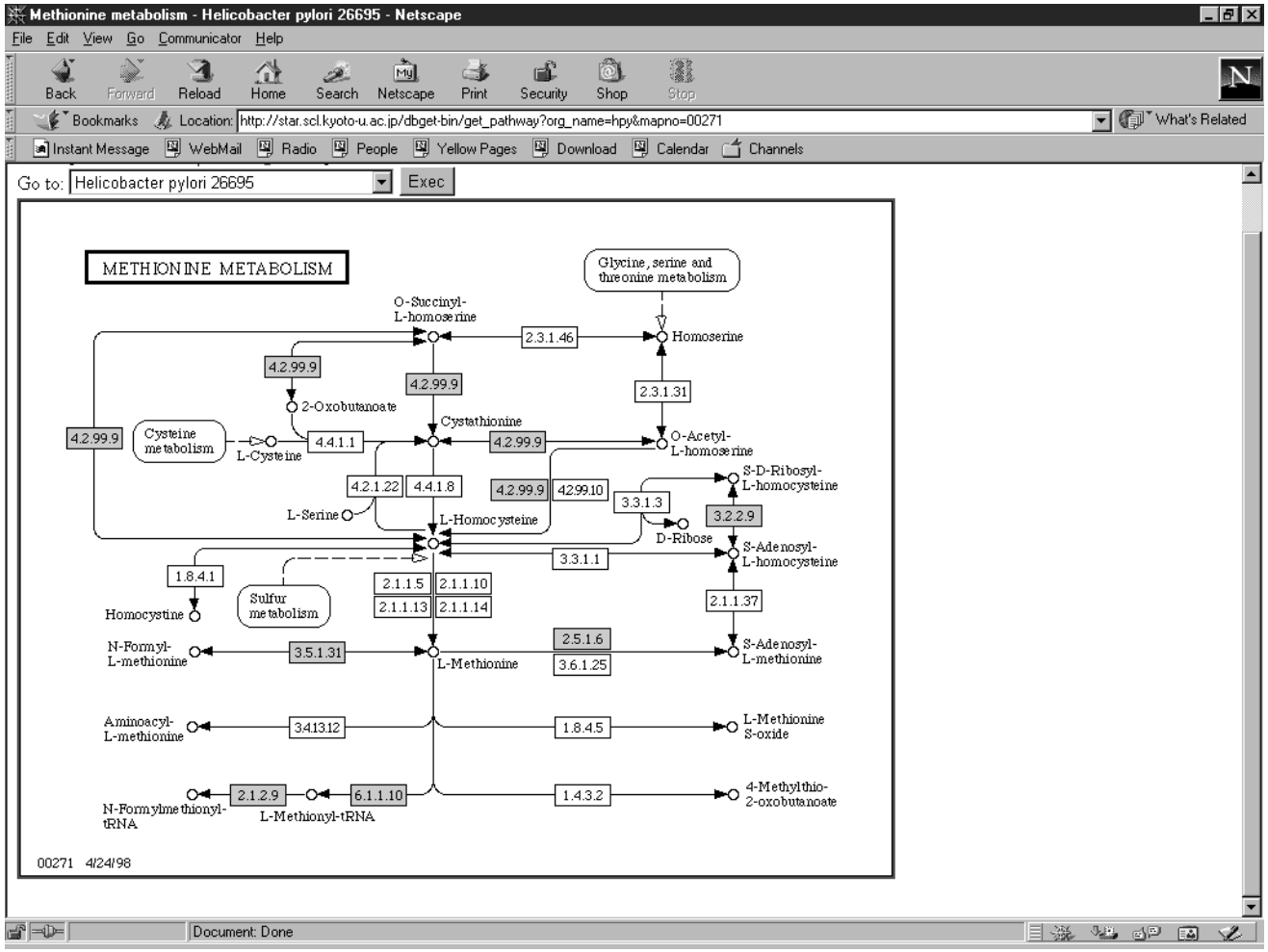


Figure 15.3. The KEGG chart of methionine metabolism in *Helicobacter pylori*. The chart shows the general organization of methionine metabolism in various species. Boxes represent enzymes with the EC numbers shown, and circles represent the intermediates of the pathway. The enzymes for which a gene has been identified in the *H. pylori* genome (according to the KEGG authors) are indicated by shaded boxes.

against all of the complete genomes available, displays color-coded functions of the detected homologs, and shows their location on a circular genome map. The output of MBGD's BLAST search also shows the degree of overlap between the query and target sequences, which could help in discerning multidomain proteins. For each sequenced genome, MBGD provides convenient lists of all recognized genes that are involved in a particular function, e.g., the biosynthesis of branched-chain amino acids or the degradation of aromatic hydrocarbons.

WIT. The WIT (''What Is There?'') database, like KEGG, aims at metabolic reconstruction for completely sequenced genomes (Overbeek et al., 2000). The distinguishing features of the WIT approach are that it (1) considers as a pathway any sequence of reactions between two bifurcations and (2) includes proteins from many partially-sequenced genomes. These features allow WIT to offer many more sequences of the same enzymes from different organisms than any other database, which significantly facilitates the recognition of additional members of these enzyme families. On the other hand, complex pathways like glycolysis or the TCA cycle have been split into many separate reactions, which sometimes makes pathway analysis unnecessarily complicated. However, anyone who overcomes the initial difficulties in using the WIT system will be rewarded by the ability to easily predict metabolic pathways in many organisms with complete and still unfinished genomes.

Organism-Specific Databases

In addition to general genomics databases, there exists a variety of databases that center around a particular organism or a group of organisms. Although all of them are useful for specific purposes, those devoted to *E. coli, B. subtilis*, and yeast are probably the ones most widely used for functional assignments in other, less studied organisms. Following are short descriptions of the most frequently used of these databases.

Escherichia coli. The importance of *E. coli* for molecular biology is reflected in the large number of databases dedicated to this organism. Two of these are maintained at the University of Wisconsin-Madison and at the Nara Institute of Technology, the research groups that carried out the actual sequencing of the *E. coli* genome. Because the Wisconsin group is now involved in sequencing the enteropathogenic *E. coli* O157:H7 and other enterobacteria, their database is most useful for analysis of enteric pathogens. The group at the Nara Institute of Technology is primarily interested in resolving the functions of still-unannotated *E. coli* genes and strives to create an ultimate resource for further studies of *E. coli*. Their site provides a convenient link of genomic data to the Kohara restriction map of *E. coli* and allows one to search for Kohara clones that cover the region of interest. Another useful database on *E. coli*, EcoCyc, lists all experimentallystudied *E. coli* genes; it also provides exhaustive coverage of the metabolic pathways identified in *E. coli*.

The goal of another *E. coli* database, EcoGene, is to provide curated sequences of the *E. coli* proteins. This is a good source for frame-shifted and potentially mistranslated proteins. Finally, Colibri and RegulonDB are the databases of choice for those interested in regulatory networks of *E. coli*. The *E. coli* Genetic Stock Center (CGSC) Web site also provides gene linkage and function information; it also lists the mutations available at the CGSC.

Mycoplasma genitalium. *Mycoplasma* has the smallest genome of all known cellular life forms, which offers some clues as to what is the lower limit of genes necessary to sustain life (the ''minimal genome''). Its comparison to the second smallest known genome, that of *Mycoplasma pneumoniae*, is available online. Recent data from TIGR provides insight into the range of *M. pneumoniae* and *M. genitalium* genes that can be mutated without loss of viability. From both computational analysis and mutagenesis studies, it appears that 250–300 genes are absolutely essential for the survival of mycoplasmas.

Bacillus subtilis. The *B. subtilis* genome also attracts considerable attention from biologists and, like that of *E. coli*, is being actively studied from the functional perspective. The Subtilist Web site, maintained at the Institute Pasteur, is constantly updated to include the most recent results on functions of new *B. subtilis* genes. In addition, a convenient index of *B. subtilis* sporulation genes is maintained at the Royal Holloway University of London.

Saccharomyces cerevisiae. The major databases specifically devoted to the functional analysis of yeast *S. cerevisiae* genome are the Saccharomyces Genome Database (SGD) at Stanford University, the Yeast Database at Munich Institute for Protein Sequences (MIPS), and Yeast Protein Database (YPD) at Proteome, Inc. All three databases provide periodically updated lists of yeast proteins with known or predicted functions, appropriate references, and mutant phenotypes and reflect the ongoing efforts aimed at complete characterization of all yeast proteins. SGD is probably the largest and most comprehensive source of information on the current status of the yeast genome analysis and includes the *Saccharomyces* Gene Registry. The MIPS database provides most of the same data and serves as a resource for new results coming from the multinational EUROFAN project. YPD is a curated database that is an useful resource for current information on the function of yeast proteins. YPD now allows free access for academic researchers using the database for non-commercial purposes.

Other useful sites for yeast genome analysis include *Saccharomyces cerevisiae* Promoter Database, listing known regulatory elements and transcriptional factors in yeast; TRansposon-Insertion Phenotypes, Localization, and Expression in *Saccharomyces* (TRIPLES) database, which tracks the expression of transposon-induced mutants and the cellular localization of transposon-tagged proteins, and the *Saccharomyces* Cell Cycle Expression Database, presenting the first results on changes in mRNA transcript levels during the yeast cell cycle.

GENOME ANALYSIS AND ANNOTATION

With recent progress in rapid, genome-scale sequencing, sequence analysis and annotation of complete genomes have become the limiting steps in most genome projects. This task is particularly daunting given the paucity of functional information for a large fraction of genes even in the best-understood model organisms, let alone poorly-studied ones such as those from Archaeal species. The standard steps involved in the structural-functional annotation of uncharacterized proteins includes (1) sequence similarity searches using programs such as BLAST, FASTA, or the Smith-Waterman algorithm; (2) identifying functional motifs and structural domains by

comparing the protein sequence against PROSITE, BLOCKS, SMART, or Pfam; (3) predicting structural features of the protein, such as likely signal peptides, transmembrane segments, coiled-coil regions, and other regions of low sequence complexity; and (4) generating a secondary (and, if possible, tertiary) structure prediction. All these steps have been automated in several software packages, such as GeneQuiz, MAGPIE, PEDANT, Imagene, and others. Of these, however, MAGPIE and PEDANT do not allow outside users to submit their own sequences for analysis and display only the authors' own results. GeneQuiz offers a limited number of searches (up to 100 a day) to general users but is still a good entry point for comparative genome analysis (Andrade et al., 1999; Hoersch et al., 2000). However, GeneQuiz relies on unrealistically high cutoff scores to infer homology, which inevitably results in relatively low sensitivity. In some cases, the user may be better off by simply using the same tools that are packaged in the aforementioned programs separately. To perform sequence analysis on a large scale, it is frequently desirable to run the requisite software locally, in batch mode. One such package that is currently available for free downloading is SEALS, developed at NCBI. It consists of a number of UNIX-based tools for retrieving sequences from GenBank, running database search programs such as BLAST and MoST, viewing and parsing search outputs, searching for sequence motifs, and predicting protein structural features (Walker and Koonin, 1997). A similar package, Imagene, has been developed at Université Paris VI (Medigue et al., 1999).

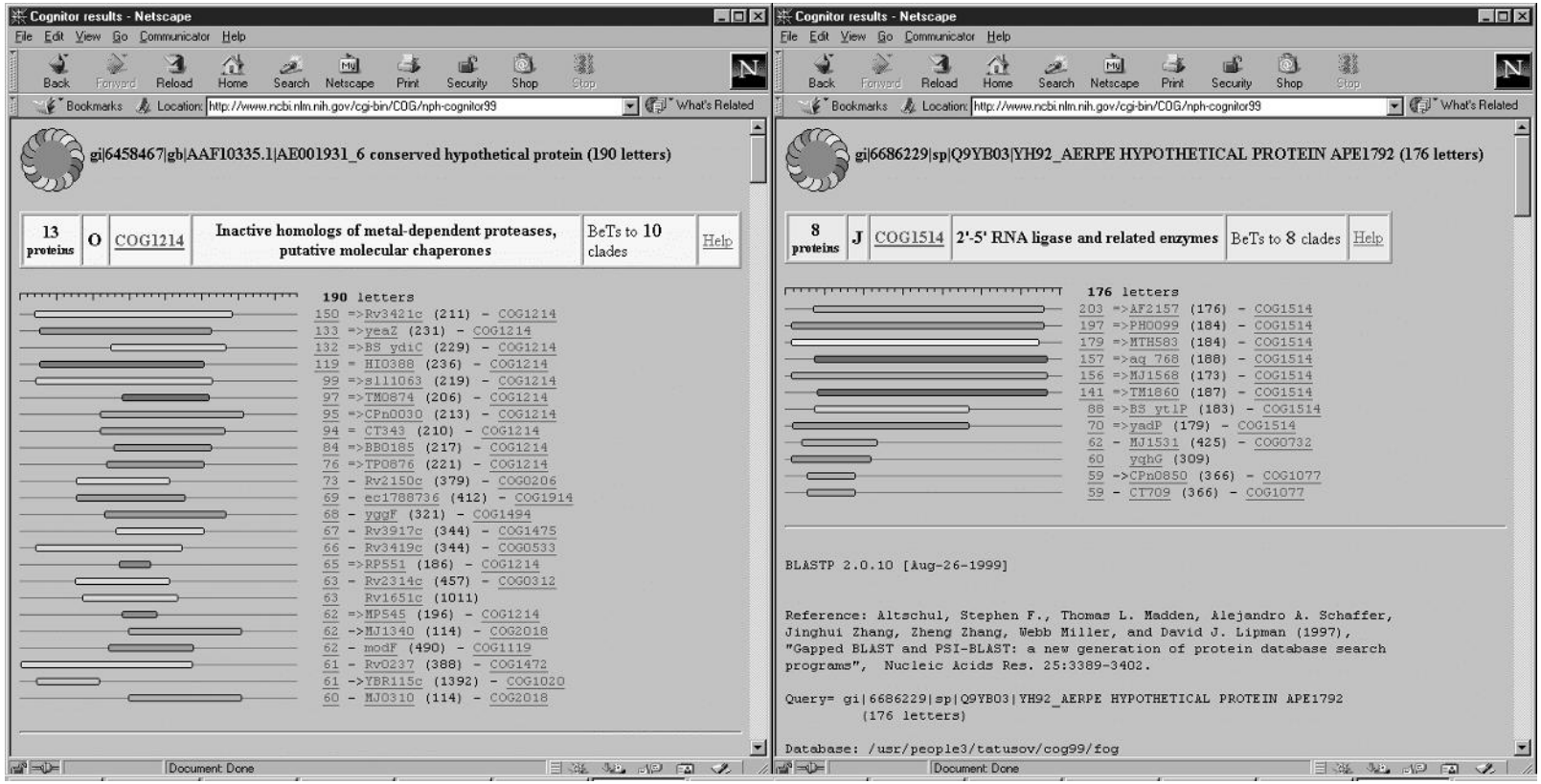
Using Genome Comparison for Prediction of Protein Functions

Analysis of the first several bacterial, archaeal, and eukaryotic genomes to be sequenced showed that the sequence comparison methods mentioned above failed to predict protein function for at least one-third of gene products in any given genome. In these cases, other approaches can be used that take into consideration all other available data, putting them into ''genome context'' (Huynen and Snel, 2000). By taking advantage of the availability of multiple complete genomes, these approaches offer new opportunities for predicting gene functions in each of these genomes. All these approaches rely on the same basic premise, that the organization of the genetic information in each particular genome reflects a long history of mutations, gene duplications, gene rearrangements, gene function divergence, and gene acquisition and loss that has produced organisms uniquely adapted to their environment and capable of regulating their metabolism in accordance with the environmental conditions. This means that cross-genome similarities can be viewed as meaningful in the *evolutionary* sense and thus are potentially useful for functional analysis. The most promising comparative methods—specifically employ information derived from multiple genomes to achieve robustness and sensitivity that are not easily attainable with standard tools. It seems that they are indeed the tools for the ''new genomics,'' whose impact will grow with the increase in the amount and diversity of genome information available. Here, some of these new approaches are briefly reviewed using for illustration, whenever possible, examples provided by currently available Web-based tools. A disproportionate number of these examples are from the COG system. This should not be construed as a claim that this is, in any sense, the best tool for genome annotation; rather, it reflects a degree of flexibility in formulating queries that is provided by the COGs as well as the subjective factor of the authors' familiarity with the organization of this system.

Transfer of Functional Information. The simplest and by far the most common way to utilize the information embedded in multiple genomes (at least at this time) is the transfer of functional information from well-characterized genomes to poorly-studied ones. Implicitly, this is done whenever a prediction is made for a newly sequenced gene on the basis of a database hit(s). There are, however, many pitfalls that tend to hamper accurate functional prediction on the basis of such hits. Perhaps the most important ones relate to the lack of sufficient sensitivity, error propagation because of reliance on incorrect or imprecise annotations already present in the general-purpose databases, and the difficulty in distinguishing orthologs from paralogs. The issue of orthology vs. paralogy is critical because transfer of functional information is likely to be reliable for orthologs (direct evolutionary counterparts) but may be quite misleading if paralogs (products of gene duplications) are involved. All these problems are, in part, obviated in the COG system, which consists of carefully annotated sets of likely orthologs and does not rely on arbitrary cutoffs for assigning new proteins to them.

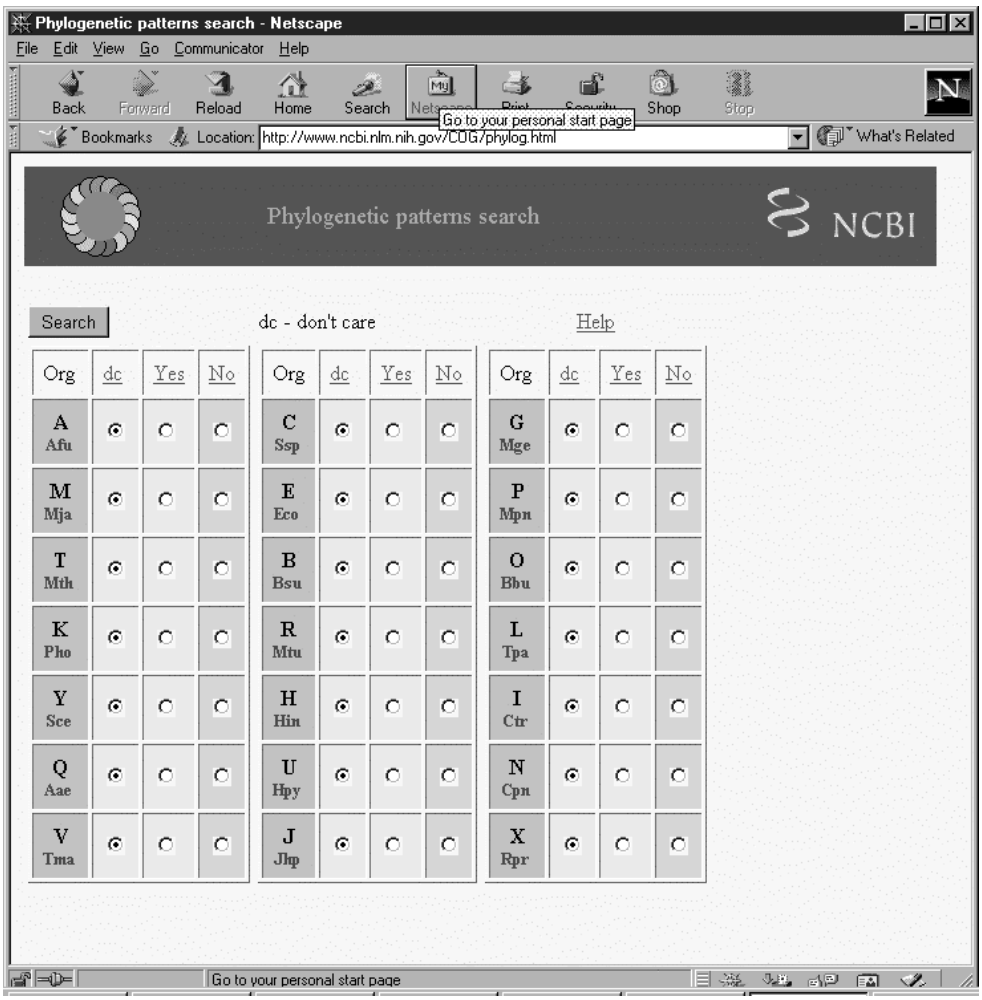
The COGs can be employed for annotation of newly-sequenced genomes using the COGNITOR program. This program assigns new proteins to COGs by comparing them to protein sequences from all genomes included in the COG database and detecting genome-specific best hits (BeTs). When three or more BeTs fall into the same COG, the query protein is considered a likely new COG member. The reasoning is that it is extremely unlikely that such coherence occurs by chance, even if the observed sequence similarity *per se* is not statistically significant. The requirement of multiple BeTs for a protein to be assigned to a COG serves, to some extent, as a safeguard against the propagation of errors that might be present in the COG database itself. Indeed, if a COG contains one or even two false-positives, this will not result in a false assignment by COGNITOR under the three-BeT cutoff rule. Figure 15.4 shows two examples of the COGNITOR application to proteins from the bacterium *Deinococcus radiodurans* and the archaeon *Aeropyrum pernix* that have not been assigned a function in the original genome annotation.

Phylogenetic Patterns (Profiles). The COG-type analysis applied to multiple genomes provides for the derivation of ***phylogenetic patterns***, which are potentially useful in many aspects of genome analysis and annotation (Tatusov et al., 1997). Similar concepts have been introduced by others in the form of phylogenetic profiles (Gaasterland and Ragan, 1998; Pellegrini et al., 1999). The phylogenetic pattern for each protein family (COG) is defined as the set of genomes in which the family is represented. The COG database is accompanied by a pattern search tool that allows the user to select COGs with a particular pattern (Fig. 15.5A). Predictably, genes that are functionally related (e.g., those that encode different subunits of the same enzyme or participate in consecutive steps of the same metabolic pathway) tend to have the same phylogenetic pattern (Fig. 15.5B). In a complementary fashion, closely related species tend to co-occur in COGs. Because of these features, phylogenetic patterns can be used to improve functional predictions in complete genomes. When a particular genome is represented in the COGs for a subset of components of a particular complex or pathway but is missing in the COGs for other components, a focused search for the latter is justified. The same applies to cases in which a gene is found in one of two closely related genomes, but not the other, particularly if it is conserved in a broad range of other genomes (Fig. 15.5C). There are several reasons why unexpectedly incomplete phylogenetic patterns may be observed.



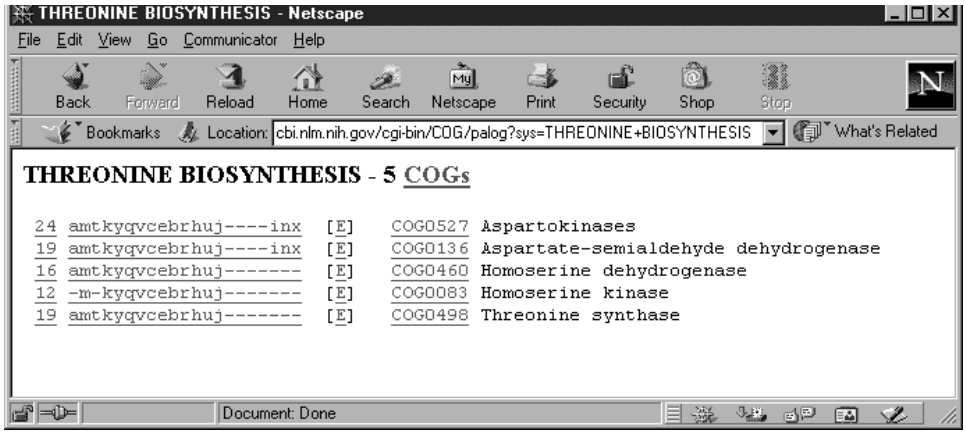
(A) (B)

Figure 15.4. Assigning proteins to COGs using COGNITOR. (A) Uncharacterized protein from the bacterium *Deinococcus radiodurans*. (B) Uncharacterized protein from the archaeon *Aeropyrum pernix*. COGNITOR compares the query sequence to the database of protein sequences from complete genomes, registers genome-specific best hits (shown by arrows), and, in case three or more of these fall within the same COG, assigns the new protein to this COG. If this happens, the COG name should be perceived as a general functional prediction for the query. For more details, see the online COG Help pages.

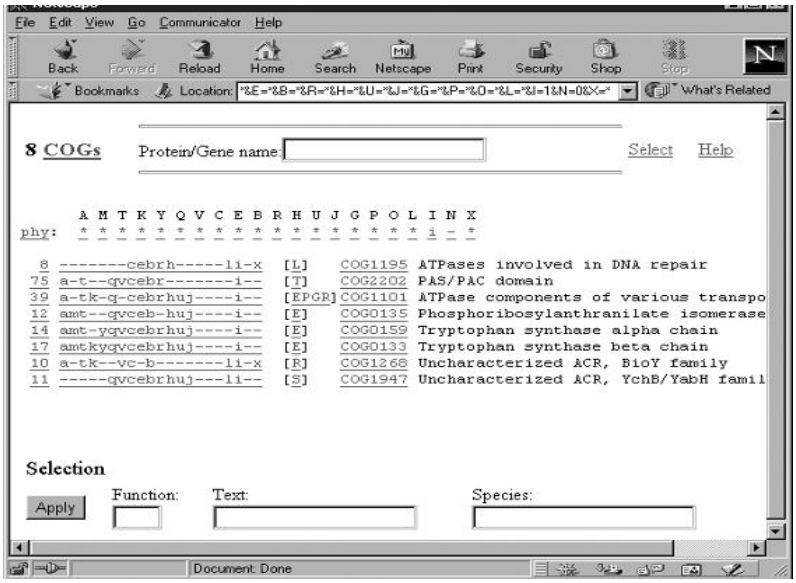


(A)

Figure 15.5. Applications of phylogenetic patterns. (A) The COG phylogenetic pattern search tool. Species name abbreviations: A (Afu), *Archaeoglobus fulgidus*; M (Mja), *Methanococcus jannaschii*; T (Mth), *Methanobacterium thermoautotrophicum*; K (Pho), *Pyrococcus horikoshii*; Y (Sce), yeast *Saccharomyces cerevisiae*; Q (Aae), *Aquifex aeolicus*; V (Tma), *Thermotoga maritima*; C (Ssp), *Synechocystis sp.*; E (Eco), *Escherichia coli*; B (Bsu), *Bacillus subtilis*; R (Mtu), *Mycobacterium tubeculosis*; H (Hin), *Haemophilus influenzae*; U (Hpy), *Helicobacter pylori*; J (Jhp), *Helicobacter pylori* J strain; G (Mge), *Mycoplasma genitalium*; P (Mpn), *Mycoplasma pneumoniae*; O (Bbu), *Borrelia burgdorferi*; L (Tpa), *Treponema pallidum*; I (Ctr), *Chlamydia trachomatis*; N (Cpn), *Chlamydia pneumoniae*; X (Rpr), *Rickettsia prowazekii*. (B) Phylogenetic pattern conservation in the threonine biosynthesis pathway. The patterns are shown using the one-letter designations for species as in (A); a letter indicates that the given species is represented in the respective COG and a dash means that it is not. Note that aspartokinase and aspartate-semialdehyde dehydrogenase are found in Chlamydiae and Rickettsiae, even though these parasitic bacteria do not encode the entire pathway of threonine biosynthesis or those for methionine and lysine biosynthesis that share the same first stages. The functions of these enzymes in these cases remain unclear. The absence of detectable homoserine kinase in *A. fulgidus* and *M. thermoautotrophicum* is probably due to nonorthologous gene displacement with a distinct kinase(s). (C) Differential genome display—COGs represented in *C. trachomatis*, but not in *C. pneumoniae*. In the generalized representation of the phylogenetic pattern above the list of COGs, asterisks indicate that the respective species may be either present or absent.



(B)



(C)

Figure 15.5. Continued

In the simplest case, certain proteins, typically small ones, could have been missed in genome translation (Natale et al., 2000). Thus, the apparent absence of *secE* genes in the genomes of *Aquifex aeolicus* and *Helicobacter pylori* that encoded all the other components of the Sec protein translocation machinery suggests that the *secE* genes could have been missed in the original genome annotation for these two bacteria. Indeed, these genes are easily recognized by searching the six-frame translation of the respective genomes using TBLASTN. Examination of the COGs that miss one representative from a group of close species similarly may result in the identification of otherwise undetected genes. For example, only one COG (COG1546) contained an *M. genitalium* protein but not an *M. pneumoniae* protein. A search for a possible missing *M. pneumoniae* counterpart identified a candidate, whose inclusion into this COG was subsequently verified by COGNITOR and sequence alignment (Natale et al., 2000).

An unexpected absence of a species in a phylogenetic pattern also may indicate that the given species encodes a highly diverged member of the respective orthologous family. For example, the presence of easily-recognizable A, B, D, and I subunits of the archaeal type H^+-ATPase in *Borrelia burgdorferi, Treponema pallidum*, and both chlamydia (COGs 1155, 1156, 1394, and 1269) immediately suggests that membrane-bound subunits of this enzyme should also be encoded in these genomes. Indeed, genes for the E and K subunits of the H^+-ATPase could be recognized in these genomes (COGs 1390 and 0636) despite their low sequence similarity to the corresponding subunits of the archaeal enzymes. The gene for the F subunit, however, has been identified so far only in *T. pallidum* but not in the three other bacterial species (see COG1436), whereas the gene for the C subunit (COG1527) has not been recognized in any of them.

Finally, unexpected ''holes'' in phylogenetic patterns and differences between components of the same complex or pathway may be the manifestation of a phenomenon termed *nonorthologous gene displacement*, in which unrelated or distantly related proteins are responsible for the same function in different organisms. When essential functions are involved, this tends to result in phylogenetic patterns that are perfectly or partially complementary, together spanning the entire range of genomes. Figure 15.6 shows two examples of COGs with such complementary phylogenetic patterns. Note that, in each case, the complementarity is not perfect because certain genomes encode both forms of the respective enzyme. In the case of lysyl-tRNA synthetases (Fig. 15.6A), the two forms of the enzyme are completely unrelated, whereas the two fructose-biphosphate aldolases (Fig. 15.6B) are distantly related, but are not orthologs. During the analysis of new genomes, it is possible to focus on families with complementary phylogenetic patterns to identify candidates for missing components of complexes and pathways.

Use of Phylogenetic Patterns for Differential Genome Display. The phylogenetic pattern approach and, specifically, the pattern search tool associated with the COGs can be used in a systematic fashion to perform formal logical operations (AND, OR, NOT) on gene sets—an approach suitably dubbed ''differential genome display'' (Huynen et al., 1997). Figure 15.7 shows examples of such analyses. This type of genome comparison allows a researcher to delineate subsets of gene products that are likely to contribute to the specific lifestyles of the respective organisms, for example, thermophily (Fig. 15.7A). The use of this approach to identify candidate drug targets in pathogenic bacteria is perhaps of special interest. It seems logical to look for such targets among those genes that are shared by several pathogenic organisms, but are missing in eukaryotes. Simple exercises in this direction show, however, that this is not a straightforward strategy. It is tempting to suggest that the best targets for new broad-spectrum antimicrobial agents would be genes that are shared by all pathogenic microbes, but not by any other organisms. The trouble is that such genes do not seem to exist, even if one allows for those that are missing in mycoplasmas, which have by far the smallest genomes (Fig. 15.7B). Furthermore, even when the conditions are relaxed so that it is only required that the genes be present in all pathogenic bacteria (except possibly mycoplasmas) and absent in yeast and *E. coli* (the dominant component of the normal gut microbial population), the net comes back empty (Fig. 15.7C). It seems therefore that the best one can do to search for such potentially universal antimicrobial agents is to isolate the genes that are present in all pathogens, possibly in other bacteria and archaea

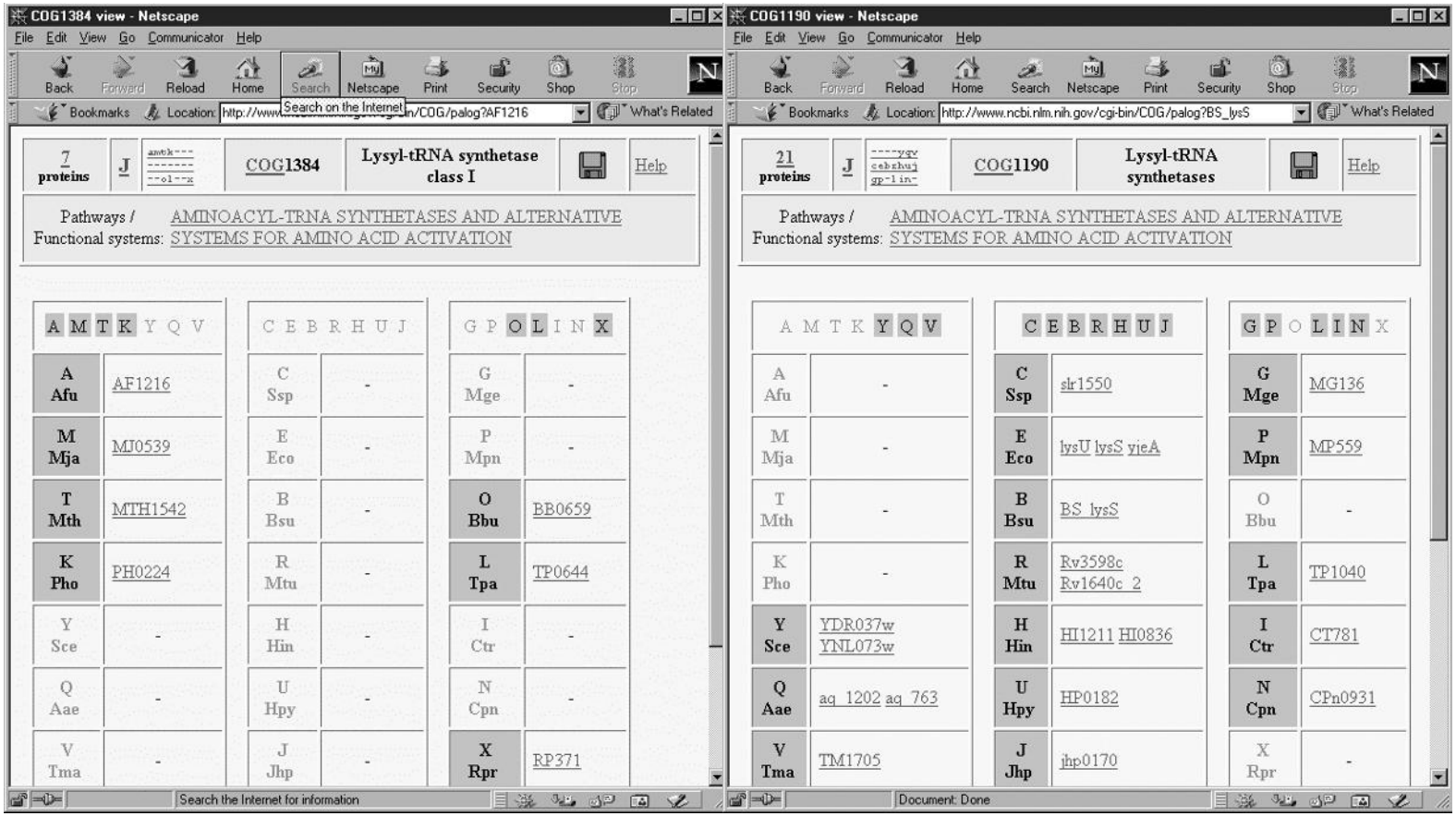
but not in eukaryotes. This results in a list of 35 families, most of which are, in fact, represented in all bacteria (Fig. 15.7D); it seems likely that some of these proteins are indeed good candidates for drug targets. More specifically directed searches can be easily set up; for example, searching for families that are represented in two species of chlamydia, possibly other pathogenic bacteria, but not any other genomes produces just two COGs (Fig. 15.7E). These could be of interest for a detailed experimental study aimed at the development of new agents that could be active against both chlamydia and spirochetes.

Examination of Gene (Domain) Fusions. Another recently developed comparative genomic approach involves systematic analysis of protein and domain fusion (and fission) (Enright et al., 1999; Marcotte et al., 1999; Snel et al., 2000). The basic assumption is that fusion would be maintained by selection only when it facilitates functional interaction between proteins, for example, kinetic coupling of consecutive enzymes in a pathway. Thus, proteins that are fused in some species can be expected to interact, perhaps physically or at least functionally, in other organisms. A straightforward example of functional inferences that can be drawn from domain fusion is seen in the histidine biosynthesis pathway, which in *E. coli* and *H. influenzae* includes two two-domain proteins, HisI and HisB (Fig. 15.8). The two domains of HisI catalyze two sequential steps of histidine biosynthesis and thus represent subunits that are likely to physically interact even when produced as separate proteins; this correlates with the predominance of the domain fusion among these enzymes (Fig. 15.8A). In contrast, the two domains of HisB catalyze the seventh and ninth steps of the pathway and hence are not likely to physically interact, which is compatible with the relative rarity of the fusion (Fig. 15.8B). The COG database includes about 700 distinct multidomain architectures that have stand-alone counterparts. Thus, using domain fusion for functional prediction has considerable heuristic potential although this approach will not work for ''promiscuous'' domains such as, for example, the DNA-binding helix-turn-helix domain, which can be found in combination with a wide variety of other domains (Marcotte et al., 1999).

In addition, several databases (with accompanying search tools) have recently been developed for detecting domains and exploring architectures of multidomain proteins: Pfam (Bateman et al., 2000), ProDom (Corpet et al., 2000), and SMART (Schultz et al., 1998, 2000).

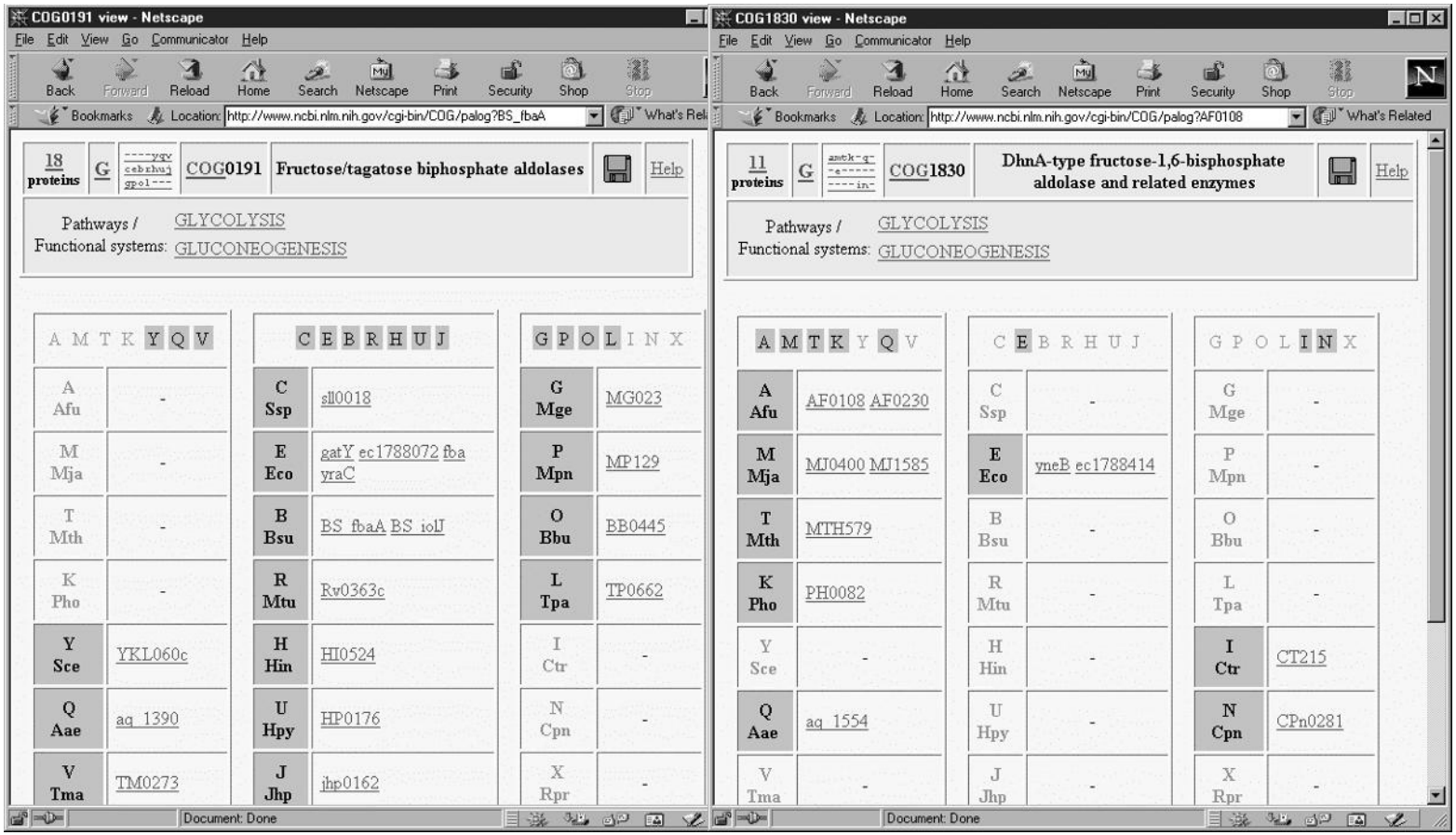
Although not comprehensive as of this writing, SMART seems to be the most advanced of these systems, combining high sensitivity of domain detection with accuracy, high speed, and extremely informative presentation of domain architectures. Rapid searches for protein domains, based on a modification of the PSI-BLAST program is now available through the Conserved Domains Database (CDD) at NCBI (cf. Chapter 11).

It seems worth considering an example of a complex multidomain protein analysis in some detail, to see how assigning functions to various domains of a multidomain protein helps one understand its likely cellular role(s). The *M. tuberculosis* protein Rv1364c consists of 653 amino acid residues. Its annotation in GenBank correctly indicates that it has statistically significant similarity to the *B. subtilis* sigma factor regulation protein, RSBU_BACSU, which, however, is only 335 amino acids long. The region of similarity between these two proteins is said to be even shorter, 244 amino acid residues. Thus, in addition to the portion apparently homologous to RsbU, Rv1364c probably contains other domains. Submitting Rv1364c for a Pfam



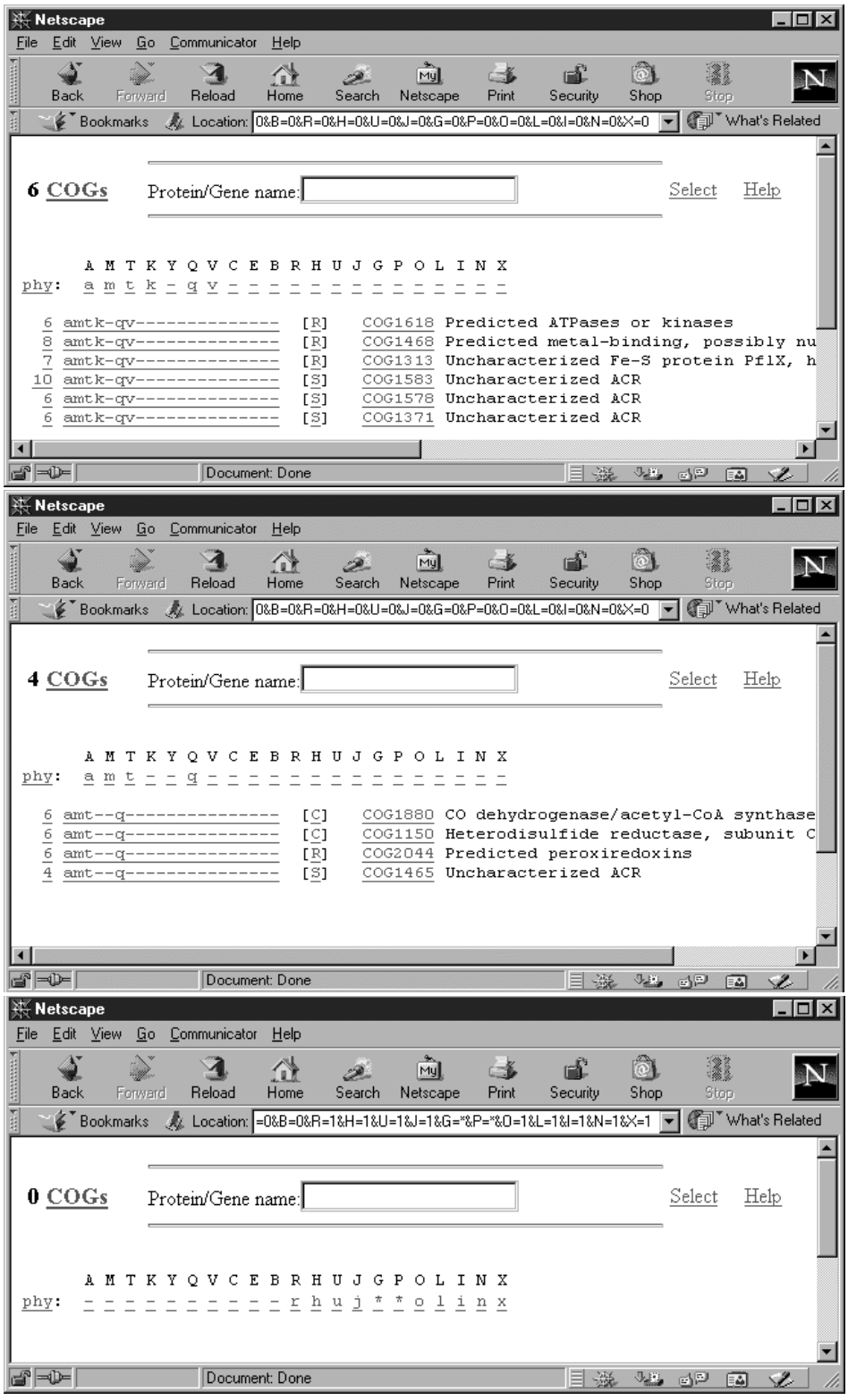
(A)

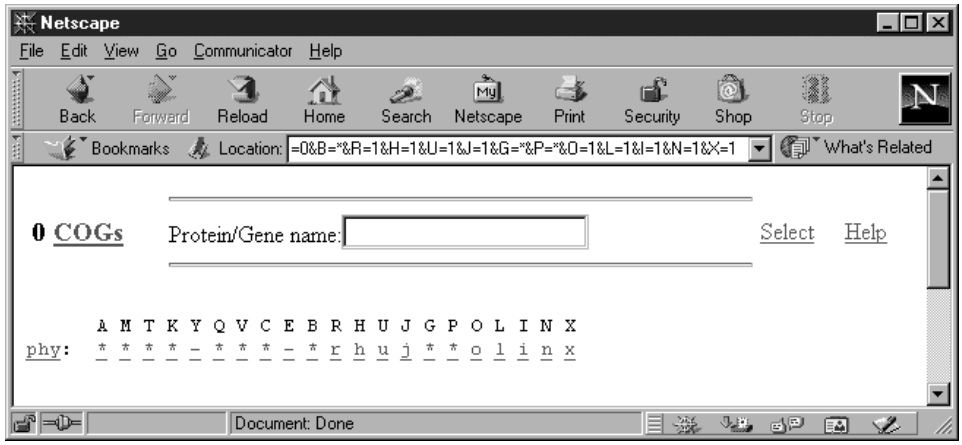
Figure 15.6. Complementary phylogenetic patterns. (A) The two classes of lysyl-tRNA synthetases. Note that both COGs include *T. pallidum*; it encodes an unusual class II enzyme that might not be involved in translation. (B) Two classes of fructose-1,6-bisphosphate aldolase. Note that *E. coli* and *A. aeolicus* encode both types of aldolases.



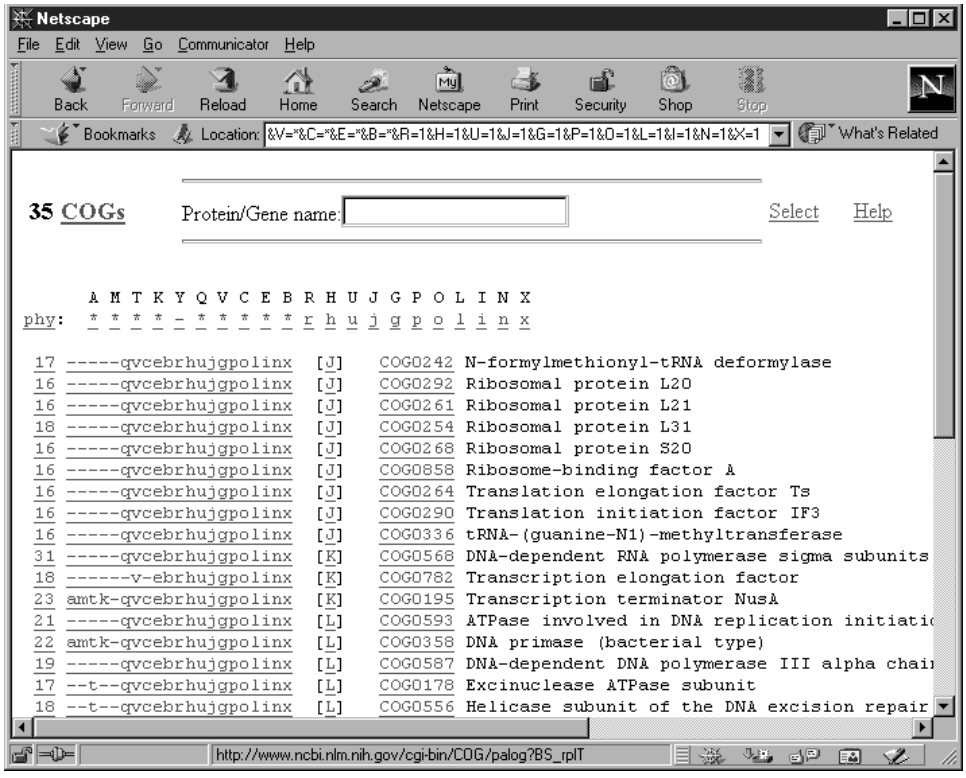
(B)

Figure 15.6. Continued


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: 0&B=0&R=0&H=0&U=0&J=0&G=0&P=0&O=0&L=0&I=0&N=0&X=0 What's Related
6 COGs Protein/Gene name: Select Help
A M T K Y Q V C E B R H U J G P O L I N X
phy: a m t k - q v - - - - - - - - - - - - - -
6 amtk-qv--------------- [R] COG1618 Predicted ATPases or kinases
8 amtk-qv--------------- [R] COG1468 Predicted metal-binding, possibly nu
7 amtk-qv--------------- [R] COG1313 Uncharacterized Fe-S protein PflX, h
10 amtk-qv--------------- [S] COG1583 Uncharacterized ACR
6 amtk-qv--------------- [S] COG1578 Uncharacterized ACR
6 amtk-qv--------------- [S] COG1371 Uncharacterized ACR
Document: Done
(A)
Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: 0&B=0&R=0&H=0&U=0&J=0&G=0&P=0&O=0&L=0&I=0&N=0&X=0 What's Related
4 COGs Protein/Gene name: Select Help
A M T K Y Q V C E B R H U J G P O L I N X
phy: a m t - - q - - - - - - - - - - - - - -
6 amt--q---------------- [C] COG1880 CO dehydrogenase/acetyl-CoA synthase
6 amt--q---------------- [C] COG1150 Heterodisulfide reductase, subunit C
6 amt--q---------------- [R] COG2044 Predicted peroxiredoxins
4 amt--q---------------- [S] COG1465 Uncharacterized ACR
Document: Done
(B)
Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: =0&B=0&R=1&H=1&U=1&J=1&G=*&P=*&O=1&L=1&I=1&N=1&X=1 What's Related
0 COGs Protein/Gene name: Select Help
A M T K Y Q V C E B R H U J G P O L I N X
phy: - - - - - - - - - - r h u j * * o l i n x
Document: Done
(C)



(D)

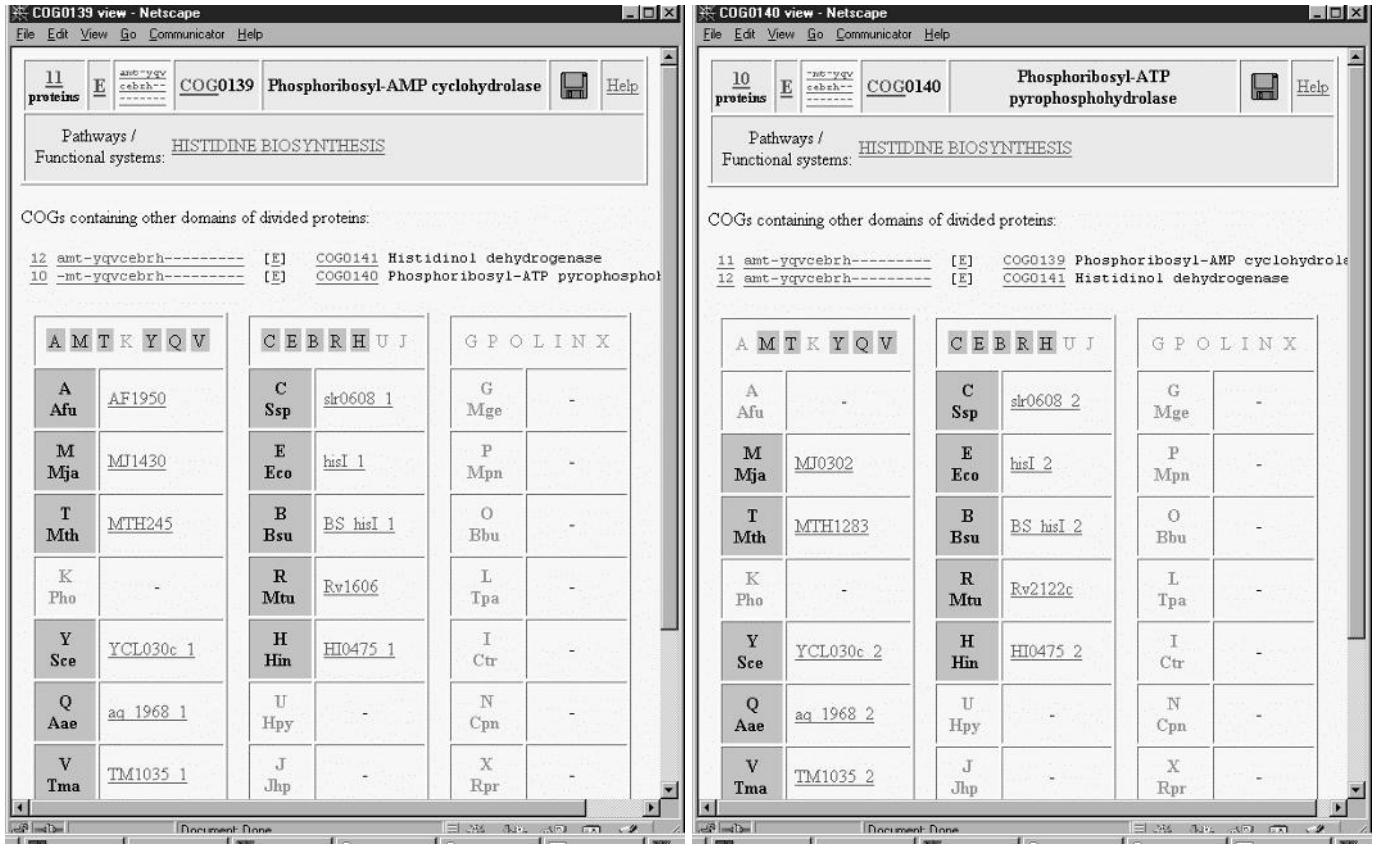


(E)

Figure 15.7. Delineation of distinct subsets of COGs using the phylogenetic pattern search tool. (A) COGs represented exclusively in thermophiles. (B) Search for COGs represented exclusively in pathogens (with a possible exception of the mycoplasmas). (C) Search for COGs represented in pathogens (with a possible exception of the mycoplasmas), but not in yeast or *E. coli*. (D) COGs represented in all pathogens, but not in yeast. (E) COGs represented in the two Chlamydia and possibly in other pathogens, but not in free-living organisms.

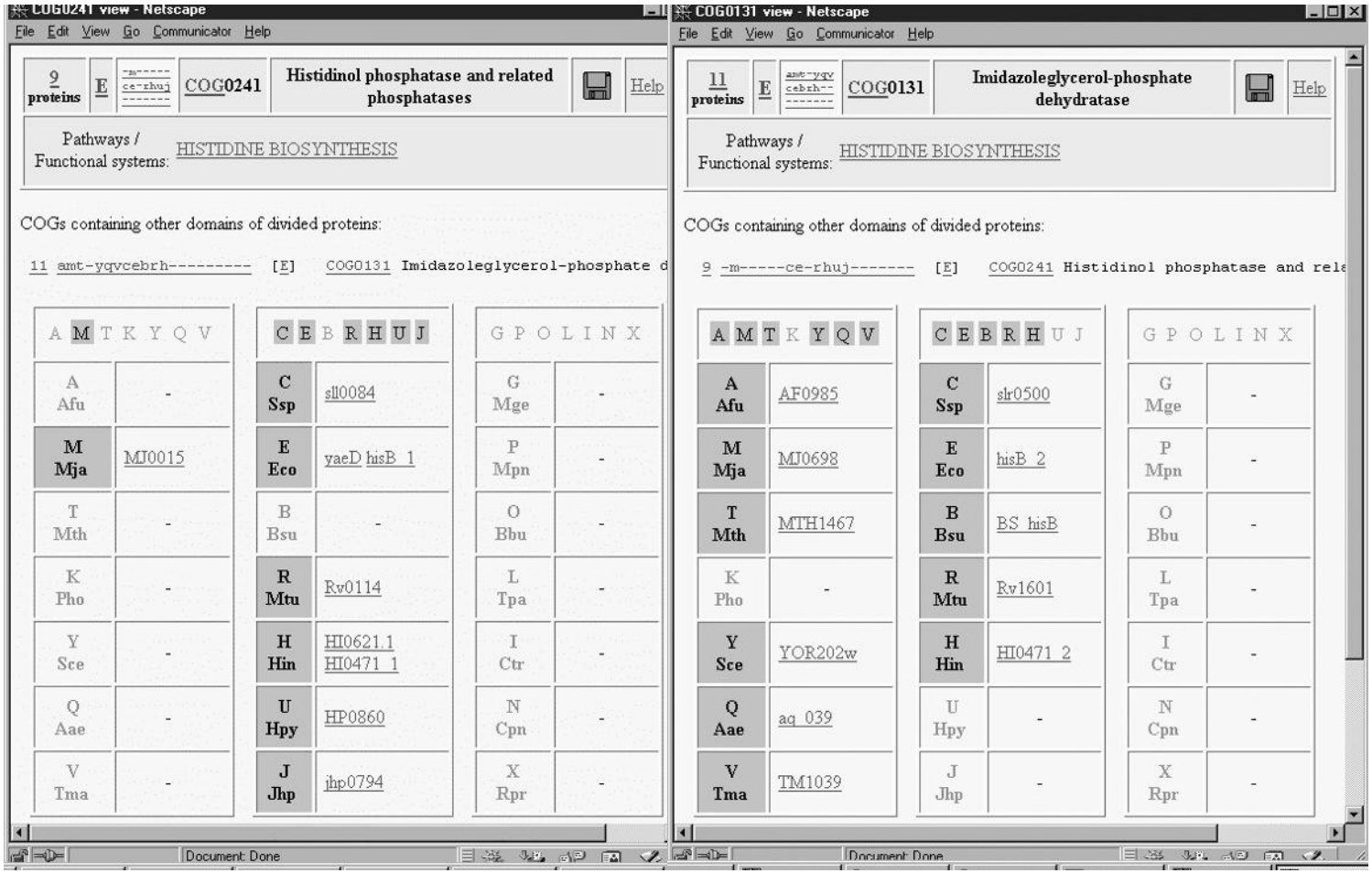
search gives an unexpected result: Pfam search identifies a SpoIIAA(RsbV)-like domain in the 550–652 region of the protein. The confidence level is not particularly high (E = 0.0049), but examination of the alignment shows conservation of the phosphorylatable serine residue and the surrounding motif, as well as conservation of the secondary structure elements (not shown). This suggests that Rv1364c actually contains *four* domains, the second and the fourth of which correspond to *B. subtilis* RsbU and RsbV proteins. The structures and functions of the first (residues 1–155) and the third (334–550) domains remain to be analyzed. This could be done by PSI-BLAST analysis of individual segments encompassing the presumptive domains, but this route would involve careful examination of the complex search outputs. In contrast, SMART provides a one-step solution (Fig. 15.9). The SMART output indicates that the protein under analysis contains N-terminal PAS and PAC domains, which are ligand-binding sensor domains present in many histidine kinases and other signal-transduction proteins, a PP2C-like phosphatase domain (the actual biochemical activity of the RsbU domain) and a histidine kinase-type ATPase domain (residues 433–527). For the latter domain, however, the statistical significance of the hit is low (E = 0.16) and the assignment needs further verification. A PSI-BLAST search started with the segment of Rv1364c identified by SMART as the ATPase domain reveals similarity to the RsbW proteins, a distinct group of serine kinases of the histidine kinase fold involved in the anti-sigma regulatory systems (hence, the low-significance hit to the general profile for this domain in SMART). SMART does not detect the C-terminal SpoIIAA domain of Rv1364c, which has been identified by Pfam, emphasizing the importance of complementary methods for complete assignments of domains and functions. The domain organization of the protein can also be probed using the COG database. Entering `Rv1364c` as a COGNITOR query assigns its domains to four COGs (where the protein already belongs since it originates from a completely sequenced genome): (1) Rv1364c_1-COG2202 ''PAS/PAC domain,'' (2) Rv1364c_2-COG2208 ''Serine phosphatase RsbU, regulator of sigma subunit,'' (3) Rv1364c_3-COG2172 ''Anti-sigma regulatory factor (Ser/Thr protein kinase),'' and (4) Rv1364c_4-COG1366 ''Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor).'' Thus, taken together, the results obtained using different methods converge on an unprecedented four-domain architecture for Rv1364c that juxtaposes the sensor PAS/PAC domain with all three components of the anti-sigma regulatory system fused within a single protein (Fig. 15.9). The PAS/PAC domain is most likely involved in sensing the energetic state of the cell, similarly to the recently characterized Aer protein of *E. coli* (Taylor and Zhulin, 1999), whereas the phosphatase, kinase, and phosphorylatable adapter domains are expected to efficiently transmit this information to the downstream signal response machinery. Thus, we can tentatively annotate Rv1364c protein as a complex regulator of sigma factor activity; the exact implications of the unusual domain fusion remain to be investigated experimentally.

Analysis of Conserved Gene Strings (Operons). An approach that is conceptually similar to the analysis of gene fusions, but is more general, if less definitive, involves systematic analysis of gene ''neighborhoods'' in genomes (Overbeek et al., 1999). Because functionally linked genes frequently form operons in bacteria and archaea, gene adjacency may provide important functional hints. Of course, many functionally related genes never form operons, and, in many instances, adjacent genes are not connected in any way. However, due to the lack of overall conservation of



(A)

Figure 15.8. Multidomain proteins in the COGs. (A) HisI proteins. Note the fusion of the two enzymes in all bacteria that possess histidine biosynthesis as opposed to the stand-alone proteins in archaea. (B) HisB protein. Note the limited phylogenetic distribution of the fusion.



(B)

Figure 15.8. Continued

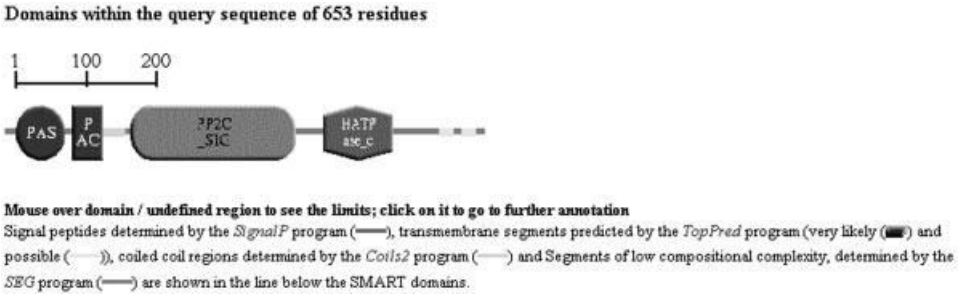


Figure 15.9. Elucidation of a protein's domain architecture using SMART. The SMART output for the Rv1364c protein is shown. The additional bar above the line shows the location of the SpoIIAA(RbsV) domain that is not recognized by SMART.

gene order in prokaryotes, the presence of a pair of adjacent orthologous genes in three or more genomes or the presence of three orthologs in a row in two genomes can be considered a statistically meaningful event and can be used to infer potential functional interaction for the products of these genes. The simplest current tool for identification of conserved gene strings in any two genomes is available as part of KEGG. It allows the user to select any two complete genomes (e.g., *B. burgdorferi* and *R. prowazekii*) and look for all genes whose products are similar to each other (e.g., have BLAST scores greater than 100) and are located within a certain distance from each other (that is, separated by 0–5 genes). The results are displayed in a graphical format illustrating the gene order and the presumed functions of gene products. In the example shown in Fig. 15.10, the uncharacterized conserved protein BB0788 from *B. burgdorferi* is similar to the RP042 protein of *R. prowazekii*, and BB0789 is similar to RP043. These pairs of genes indeed have been identified as orthologs in the COGs (COG0037 and COG0465, respectively), and examination of the relative genomic locations of other members of these COGs shows that orthologous gene strings are present in the genomes of *C. trachomatis* (CT840-CT841), *C. pneumoniae* (CPn0997-CPn0998), and *T. maritima* (TM0579-TM0580), whereas in *B. subtilis* the corresponding genes (*yacA* and *ftsH*) are one gene apart. This conservation of gene juxtaposition in phylogenetically distant bacteria is suggestive of a functional connection. The functions of one of the genes in all these pairs are well known, as they are clear orthologs of the *E. coli ftsH* (*hflB*) gene. This gene encodes an ATP-dependent metalloprotease, which is responsible for the degradation of short-lived cytoplasmic proteins and a distinct class of membrane proteins. By association, BB0788 and its orthologs may be expected to also play some, perhaps regulatory, role in the degradation of specific protein classes. The *E. coli* ortholog of BB0788, MesJ, is annotated in the SWISS-PROT database as a putative cell cycle protein. We have been unable to identify the source of this information. Nevertheless, it is compatible with a functional (and possibly also physical) interaction between MesJ and FtsH, which also has been implicated in cell division on the basis of genetic data. In the COG database, MesJ and its orthologs (COG0037) are annotated as ''Predicted ATPases of the PP-loop superfamily.'' All these enzymes share a diagnostic sequence motif, which has been discovered previously in a number of ATP pyrophosphatases (Bork and Koonin, 1994). This motif is clearly seen in the multiple alignment accompanying the COG and in the ProDom entry PD000352. Therefore,

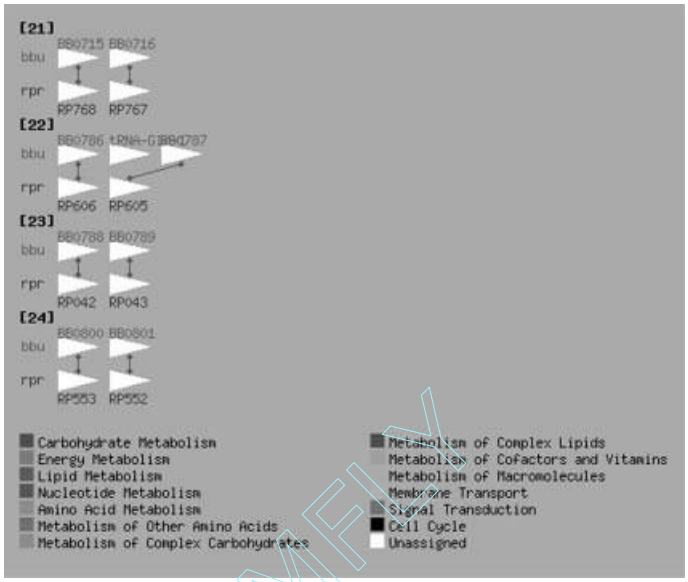


Figure 15.10. The results of a search for conserved gene strings (operons) in *Borrelia burgdorferi* (upper arrows) and *Rickettsia prowazekii* (lower arrows) using the KEGG gene cluster tool. The arrows are color-coded according to the gene function. Genes without an assigned function (all of the genes in this page) are shown in white.

by combining operon information with sequence-based prediction of the biochemical activity, we hypothesize that MesJ and its orthologs are ATP-pyrophosphatases involved in the regulation of FtsH-mediated proteolysis of specific bacterial proteins, which may be important for cell division. Because a PP-loop superfamily ATPase would comprise a novel class of cell division regulators, experimental verification of this hypothesis will be of considerable interest.

APPLICATION OF COMPARATIVE GENOMICS—RECONSTRUCTION OF METABOLIC PATHWAYS

To succinctly recap the genome analysis tools discussed above, we present here a reconstruction of the glycolytic pathway in the archaeon *Methanococcus jannaschii*. Metabolic reconstruction is one of the indispensable final steps of all genome analyses and a natural convergence point for the data produced by different methods. Glycolysis is perhaps *the* central pathway of cellular biochemistry as it becomes obvious from a cursory exploration of the general scheme of biochemical pathways, available in the interactive form on the ExPASy Web site. The names of all the enzymes and metabolites on this map are hyperlinked and searchable. Entering ''glycolysis'' as the search term finds three fields, B5, F5 and U6, the first two of which indicate the border areas of glycolysis, which actually stretches from C5 to E5. The enzyme names are hyperlinked to the ENZYME database, which is now the official site of the Enzyme Commission (IUPAC-IUBMB Joint Commission for Biochemical Nomenclature). The ENZYME database lists names and catalyzed reactions for all

the enzymes that have been assigned official Enzyme Commission (EC) numbers, whether or not their protein sequences are known. Thus, clicking on the name ''phosphoglucomutase'' will bring up the corresponding page in the ENZYME database, which will also indicate that the official name of this enzyme is glucose-6-phosphate isomerase (EC 5.3.1.9).

Glycolysis Step-by-Step

1. Glucose-6-Phosphate Isomerase. Each page in the ENZYME database lists all enzymes with the corresponding EC number that are included in the current release of SWISS-PROT. There is, however, no entry for *M. jannaschii* under glucose-6-phosphate isomerase; therefore, we will instead use the KEGG database. On entering KEGG, one can go to Open KEGG, then to Metabolic Pathways, then to Glycolysis. This takes one to an easy-to-navigate chart of compounds and enzymes that participate in glycolysis and gluconeogenesis. The pull-down menu in the upper left corner allows one to select the organism of choice. When *Methanococcus jannaschii* is selected, the boxes that indicate enzymes already recognized in this organism become shaded green. One can see that the box 5.3.1.9 is highlighted. Clicking on this box shows *M. jannaschii* protein MJ1605, which indeed can be confidently identified as phosphoglucomutase (glucose-6-phosphate isomerase). In the SWISS-PROT/TREMBL database, this protein (Q59000) is currently annotated as ''similar to prokariota glucose-6-phosphate isomerase.'' To verify that MJ1605 is indeed the ortholog of known glucose-6-phosphate isomerases, one can use the COGs, WIT, or MBGD systems. For example, in the COG database, MJ1605 is the only member of COG0166 from *M. jannaschii*; since this COG includes glucose-6-phosphate isomerases from a number of species, identification of the *M. jannaschii* protein appears to be reliable. This can be confirmed by examination of the multiple sequence alignment associated with this COG, which shows a particularly high similarity between MJ1605 and phosphoglucoisomerases from *Thermotoga maritima* and *B. subtilis*. The WIT database will show that MJ1605 is even more similar to the (predicted) glucose-6-phosphate isomerases of *Campylobacter jejuni* and *Streptococcus pyogenes*. Collectively, this evidence leaves no doubt that we have identified the correct *M. jannaschii* protein.

2. Phosphofructokinase. The next glycolytic enzyme, phosphofructokinase (EC 2.7.1.11), illustrates the opportunities and limitations of metabolic reconstruction based on comparative genomics. The most common version of this enzyme, PfkA, uses either ATP (in bacteria and many eukaryotes) or pyrophosphate (primarily in plants) as the phosphate donor. In addition, *E. coli* encodes a second version of this enzyme, PfkB, which is unrelated to PfkA and instead belongs to the ribokinase family of carbohydrate kinases. All the databases we can use agree that there is no readily identifiable candidate for this activity in *M. jannaschii*. Indeed, the KEGG chart for *M. jannaschii* does not show this enzyme as predicted, WIT does not suggest any candidates for this function, and the COG database does not show any archaeal members in its COG0205 (6-phosphofructokinase) or COG1105 [fructose-1-phosphate kinase and related fructose-6-phosphate kinase (PfkB)] entries. Thermophilic archaea possess a distinct, ADP-dependent phosphofructokinase, the gene(s) for which has been recently identified in *Pyrococcus furiosus* (Tuininga et al., 1999).

An ortholog of this protein is readily identifiable in other Pyrococci and in *M. jannaschii* (MJ1604) but not in any other archaeal or bacterial species; its sequence shows no detectable similarity to other kinases. This is a clear case of nonorthologous gene displacement; due to the limited phylogenetic distribution of this novel Pfk, the candidate protein in *M. jannaschii* could not have been detected by computational means until its ortholog had been experimentally characterized in *P. furiosus.*

3. Aldolase. The next glycolytic enzyme, fructose-1,6-bisphosphate aldolase (EC 4.1.2.13), is found in two substantially different variants, namely, metal-independent (class I) and metal-dependent (class II) aldolases in bacteria and eukaryotes. There is no ortholog of either of these in *M. jannaschii*. Instead, predicted archaeal aldolases comprise COG1830 (''DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes''), which includes two *M. jannaschii* proteins, MJ0400 and MJ1585. This COG includes orthologs of the recently described class I aldolase of *E. coli*, which is only very distantly related to the typical class I enzymes (Galperin et al., 2000). As indicated above, the phylogenetic patterns for typical bacterial class II aldolase (COG0191) and this DhnA-type aldolase (COG1830) complement each other, with the exception that both types of aldolases are encoded by *E. coli* and *A. aeolicus*. This complementarity allows one to predict that the DhnA-type enzyme functions as the only fructose-1,6-bisphosphate aldolase in chlamydiae and archaea, including *M. jannaschii*.

4–6. Triosephosphate Isomerase, Glyceraldehyde-3-Phosphate Dehydrogenase, and Phosphoglycerate Kinase. These enzymes catalyze the next three steps of glycolysis and are nearly-uniformly represented in all organisms. The *M. jannaschii* candidates for these activities (MJ1528, MJ1146, and MJ0641, respectively) can be easily identified by a BLAST search. Accordingly, all major databases converge on these functional assignments.

7. Phosphoglycerate Mutase. The activity of phosphoglycerate mutase (EC 5.4.2.1) has been experimentally demonstrated in a close relative of *M. jannaschii*, but there are no obvious candidate proteins to carry out this function. As a result, KEGG does not show this enzyme as encoded in the *M. jannaschii* genome. WIT does suggest a candidate protein (MJ1612 or RMJ05975 in WIT) but annotates it as ''phosphonopyruvate decarboxylase.'' Indeed, WIT shows only limited sequence similarity of this protein to the phosphoglycerate mutases from the mycoplasmas and *H. pylori*, whereas all close homologs of MJ1612 seem to be phosphonopyruvate decarboxylases. Finally, a search of the COG database for ''phosphoglycerate mutase'' retrieves three COGs, only one of which, COG1015 (phosphopentomutase/predicted phosphoglycerate mutase and related enzymes), contains archaeal members, including two *M. jannaschii* proteins, MJ1612 and MJ0010. A detailed analysis shows that both of them could possess phosphoglycerate mutase activity (Galperin et al., 1998). The definitive identification of the phosphoglycerate mutase in *M. jannaschii* awaits direct biochemical studies.

8–9. Enolase and Pyruvate Kinase. In *M. jannaschii*, these enzymes are readily identified through strong sequence similarity to the corresponding bacterial orthologs. As a result, there is general consensus on assigning these functions to MJ0232 and MJ0108, respectively. Thus, although no glycolytic enzymes has been

experimentally characterized in *M. jannaschii*, computational analysis provides for the identification of all of them, with uncertainty remaining with regard to just one step.

Glycolysis in* H. pylori? *A Cautionary Note. Metabolic reconstruction based on genome analysis requires considerable caution to be exerted with regard to the plausibility of functional predictions in the general context of the biology of the organism in question. So as to not depart from glycolysis, consider the case of phosphofructokinase in the gastric ulcer-causing bacterium, *Helicobacter pylori*. This organism lacks a homolog of PfkA but does encode a close homolog of PfkB. Accordingly, WIT suggests this PfkB homolog, HP0858, as a candidate for the phosphofructokinase function in *H. pylori*. Perusal of the COG database, however, suggests a different solution, since HP0858 consists of two distinct domains, one of which indeed belongs with PfkB and other sugar kinases (COG0524), whereas the other one is predicted to be a nucleotidylyltransferase (COG0615). Given this domain architecture, it appears most likely that this protein is an ADP-heptose synthetase, an enzyme of peptidoglycan biosynthesis. A simple analysis of the biology of *H. pylori* as an acid-tolerant bacterium shows that it is likely to use gluconeogenesis, but not glycolysis, which makes phosphofructokinase unnecessary for this organism. Thus, the assignment of this activity to HP0858, which appeared to be statistically supported, is, in all likelihood, biologically irrelevant. Examination of this protein's domain architecture could be the first indication of its role in a process other than glycolysis, with biological considerations further supporting this interpretation.

AVOIDING COMMON PROBLEMS IN GENOME ANNOTATION

Due to its intrinsic complexity, genome annotation defies full automation and is inherently errorprone. Accidental error rate can be minimized only through further development of the semiautomated annotation systems and the appropriate training of annotators. There are, however, several sources of systematic error that plague genome analysis. Awareness of these could help improve the quality of genome annotation (Brenner, 1999; Galperin and Koonin, 1998).

Error Propagation and Incomplete Information in Databases

Sequence databases are prone to error propagation, whereby erroneous annotation of one protein causes multiple errors as it is used for annotation of new genomes. Furthermore, database searches have the potential for noise amplification, so that the original annotation could have involved a minor inaccuracy or incompleteness, but its transfer on the basis of sequence similarity aggravates the problem and eventually results in outright false functional assignments (Bhatia et al., 1997). These aspects of current sequence databases make the common practice of assigning gene function on the basis of the annotation of the best database hit (or even a group of hits with compatible annotations) highly error-prone. Time- and labor-consuming as this may be, adequate genome annotation requires that each gene be considered in the context of both its phylogenetic relationships and the biology of the respective organism, hence the rather disappointing performance of automated systems for genome annotation. There are numerous reasons why functional annotation may be

wrong in the first place, but two main groups of problems have to do with database search methods and with the complexity and diversity of the genomes themselves.

False Positives and False Negatives in Database Searches

It is customary in genome annotation to use a cutoff for ''statistically significant'' database hits. It can be expressed in terms of the false-positive expectation (E) value for the BLAST searches and is set routinely at values such as $E = 0.001$ or $E = 10^{-5}$. The problem with this approach is that the distribution of similarity scores for evolutionarily and functionally relevant sequence alignments is very broad and that a considerable fraction of them fail the E-value cutoff, resulting in undetected relationships and missed opportunities for functional prediction (false negatives). Conversely, spurious hits may have E-values lower than the cutoff, resulting in false positives. The latter is most frequently caused by compositional bias (low-complexity regions) in the query sequence and in the database sequences. Clearly, there is a trade-off between *sensitivity* (false-negative rate) and *selectivity* (false-positive rate) in all database searches, and it is particularly difficult to optimize the process in genome-wide analyses. There is no single recipe to circumvent these problems. To minimize the false-positive rate, appropriate procedures for filtering low-complexity sequences are critical (Wootton and Federhen, 1996). Filtering using the SEG program is the default for Web-based BLAST searches, but additional filtering is justified for certain types of proteins. For example, filtering of predicted nonglobular domains using SEG with specifically adjusted parameters and filtering for coiled-coil domains using the COILS2 program is one way to minimize the false positive rate. Minimizing the false-negative rate (that is, maximizing sensitivity) is an open-ended problem. It should be kept in mind that a standard database search (e.g., using BLAST) with the protein sequences encoded in the given genome as queries is insufficient for an adequate annotation. To increase the sensitivity of genome analysis, it should be supplemented by other, more powerful methods such as screening the set of protein sequences from the given genome with preformed profile libraries (see above).

Genome, Protein, and Organismal Context as a Source of Errors

As discussed above, protein domain architecture, genomic context and an organism's biology may serve as sources of important, even if indirect, functional information. However, those same context features, if misinterpreted, may become one of the major sources of error and confusion in genome annotation. Standard database search programs are not equipped with the means to explicitly address the implications of the multidomain organization of proteins. Therefore, unless specialized tools such as SMART or COGs are employed and/or the search output is carefully examined, assignment of the function of a single-domain protein to a multidomain homolog and vice versa becomes frequent in genome annotation. Promiscuous, mobile domains are particularly likely to wreak havoc in the annotation process, as demonstrated, for example, by the proliferation of ''IMP-dehydrogenase-related proteins'' in several genomes. In reality, most or all of these proteins (depending on the genome) share with IMP dehydrogenase the mobile CBS domain but not the enzymatic part (Galperin and Koonin, 1998).

As discussed above, it is also critical for reliable genome annotation that the biological context of the given organism is taken into account. In a simplistic ex-

ample, it is undesirable to annotate archaeal gene products as nucleolar proteins, even if their eukaryotic homologs are correctly described as such. As a general guide to functional annotation, it should be kept in mind that current methods for genome analysis, even the most powerful and sophisticated of them, facilitate, but do not supplant the work of a human expert.

CONCLUSIONS

With an increasing number of complete genome sequences becoming available and specialized tools for genome comparison being developed, the comparative approach is becoming the most powerful strategy for genome analysis. It seems that the future should belong to databases and tools that consistently organize the genomic data according to phylogenetic, functional, or structural principles and explicitly take advantage of the diversity of genomes to increase the resolution power and robustness of the analysis. Many tasks in genome analysis can be automated, and, given the rapidly growing amount of data, automation is critical for the progress of genomics. This being said, the ultimate success of comparative genome analysis and annotation critically depends on complex decisions based on a variety of inputs, including the unique biology of each organism. Therefore, the process of genome analysis and annotation taken as a whole is, at least at this time, not automatable, and human expertise is necessary for avoiding errors and extracting the maximum possible information from the genome sequences.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 15

GENERAL	
NCBI	*http://www.ncbi.nlm.nih.gov/*
EBI	*http://www.ebi.ac.uk/*
DDBJ	*http://www.ddbj.nig.ac.jp/*
ExPASy	*http://www.expasy.ch/*
GENOME PROJECTS	
GenBank Entrez Genomes division	*http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html*
The Institute for Genome Research (TIGR) Microbial Database	*http://www.tigr.org/tdb/mdb/mdb.html*
Integrated Genomics Inc.	*http://wit.integratedgenomics.com/GOLD*
NHGRI List of Genetic and Genomic Resources	*http://www.nhgri.nih.gov/Data*
The Sanger Centre	*http://www.sanger.ac.uk*
Washington University-St. Louis	*http://genome.wustl.edu*
Ohlahoma University	*http://www.genome.ou.edu/*
Microbial Genome Database	*http://mbgd.genome.ad.jp*

GENOME ANALYSIS SYSTEMS	
MAGPIE	*http://genomes.rockefeller.edu/magpie*
GeneQuiz	*http://jura.ebi.ac.uk:8765/ext-genequiz/*
PEDANT	*http://pedant.mips.biochem.mpg.de*
Clusters of Orthologous Groups of Proteins (COGs)	*http://www.ncbi.nlm.nih.gov/COG*
Kyoto Encyclopedia of Genes and Genomes (KEGG)	*http://www.genome.ad.jp/kegg*
What Is There (WIT)	*http://wit.integratedgenomics.com/IGwit/*
DATABASES AND TOOLS FOR ANALYSIS OF PROTEIN DOMAINS	
ProDom	*http://protein.toulouse.inra.fr/prodom.html*
Pfam	*http://pfam.wustl.edu*
SMART	*http://smart.embl-heidelberg.de*
Protein modules	*http://www.bork.embl-heidelberg.de/Modules/sinput.shtml*
CDD search	*http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi*
INDIVIDUAL MICROBIAL GENOME DATABASES	
Escherichia coli	
University of Wisconsin-Madison	*http://www.genetics.wisc.edu*
Nara Institute of Technology	*http://ecoli.aist-nara.ac.jp*
EcoCyc	*http://ecocyc.doubletwist.com*
EcoGene	*http://bmb.med.miami.edu/ecogene*
Colibri	*http://bioweb.pasteur.fr/GenoList/Colibri*
RegulonDB	*http://www.cifn.unam.mx/Computational_Biology/regulondb*
E. coli Genetic Stock Center (CGSC)	*http://cgsc.biology.yale.edu*
Mycoplasma genitalium	
Mycoplasma genome	*http://www.zmbh.uni-heidelberg.de/M_pneumoniae/*
Essential genes	*http://www.sciencemag.org/feature/data/1042937.shl*
Bacillus subtilis	
Subtilist	*http://bioweb.pasteur.fr/GenoList/SubtiList*
Sporulation genes	*http://www1.rhbnc.ac.uk/biological-sciences/cutting*
Yeast	
Saccharomyces Genome Database (SGD)	*http://genome-www.stanford.edu/Saccharomyces*
MIPS Yeast Database	*http://www.mips.biochem.mpg.de/proj/yeast*
Yeast Protein Database (YPD)	*http://www.proteome.com/databases*
Promoter Database	*http://cgsigma.cshl.org/jian*
TRIPLES database	*http://ygac.med.yale.edu*
Cell Cycle Expression Database	*http://genomics.stanford.edu*

Metabolic reconstruction

Biochemical pathways map	*http://www.expasy.ch/cgi-bin/search-biochem-index*
ENZYME	*http://www.expasy.ch/enzyme/*

PROBLEMS FOR ADDITIONAL STUDY

The reader who decides to address these problems is expected to use the tools described in this chapter including general-purpose ones such as different versions of BLAST. In most cases, it is advisable to apply more than one method when trying to solve a problem. The reader should keep in mind that some of the problems may not have a single ''correct'' solution but rather one or perhaps even two or three most likely solutions.

1. Three archaeal species, *M. jannaschii, M. thermoautotrophicum*, and *A. fulgidus*, typically are found in COGs together. The fourth species, *P. horikoshii*, is frequently missing from these COGs (see the table of co-occurrence of genomes in COGs, which is available on the COG Web site). How do you explain this? What kind of genes are absent from *P. horikoshii*?

2. Which bacterial species share the greatest number of genes with Archaea, if measured as the ratio of the shared genes (COGs) to the total number of genes in the bacterial genome? Use the table of co-occurrence of genomes in COGs to identify the trend. Once it is clear, discuss different explanations for the observations.

3. What is the function of the *E. coli* HemK protein and its orthologs? Explain the basis for your conclusion and possible alternatives.

4. How many IMP dehydrogenases are there in *A. fulgidus*? In *A. aeolicus*? What are the domain organizations and functions of ''IMP dehydrogenase-related'' proteins?

5. Methanobacterial protein MTH1425 is annotated in GenBank as *O*-sialoglycoprotein endopeptidase. Describe the domain organization of this protein. Should you detect an unexpected domain fusion, could it be due to a sequencing error? What are the possible functional implications?

6. What are the functions of the following proteins:

a) MJ1612
b) MJ1001
c) *E. coli* NagD

7. Describe the domain architectures and functions and of the following proteins:

a) *E. coli* YfiQ
b) *B. subtilis* YtvA
c) slr1759

8. Suggest a comparative-genomic approach to search for new targets for anti-ulcer drugs; which of the identified proteins could be also potential targets for anti-tuberculosis drugs?

9. The Glu-tRNA^{Gln} amidotransferase consists of A, B, and C subunits. Compare the phylogenetic patterns for these. How can you explain the differences? Test some of the explanations. What is unusual about the Glu-tRNA^{Gln} amidotransferase complex of the mycoplasmas?

10. Families of orthologous proteins involved in translation frequently contain one representative of each of the bacterial and archaeal genomes, but two members from yeast. How would you explain this pattern? Use at least two lines of evidence in support of your explanation.

REFERENCES

Andrade, M. A., Brown, N. P., Leroy, C., et al. (1999). Automated genome sequence analysis and annotation. *Bioinformatics*. 15, 391–412.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The pfam protein families database. *Nucleic Acids Res.* 28, 263–266.

Bhatia, U., Robison, K., and Gilbert, W. (1997). Dealing with database explosion: a cautionary note. *Science*. 276, 1724–1725.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*. 277, 1453–1474.

Bork, P., and Koonin, E. V. (1994). A P-loop-like motif in a widespread ATP pyrophosphatase domain: implications for the evolution of sequence motifs and enzyme activity. *Proteins*. 20, 347–355.

Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* 15, 132–133.

Bult, C. J., White, O., Olsen, G. J., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science.* 273, 1058–1073.

Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 28, 267–269.

Enright, A. J., Illopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature.* 402, 86–90.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113.

Fleischmann, R. D., Adams, M. D., White, O., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* 269, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., et al. (1995). The minimal gene complement of *Mycoplasma genitalium. Science*. 270, 397–403.

Frishman, D., and Mewes, H.-W. (1997). PEDANTic genome analysis. *Trends Genet.* 13, 415–416.

Gaasterland, T., and Ragan, M. A. (1998). Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics*. 3, 199–217.

Galperin, M. Y., Aravind, L., and Koonin, E. V. (2000). Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in archaea. *FEMS Microbiol. Lett.* 183, 259–264.

Galperin, M. Y., Bairoch, A., and Koonin, E. V. (1998). A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases. *Protein Sci.* 7, 1829–1835.

Galperin, M. Y., and Koonin, E. V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *InSilico Biol*. 1, 55–67.

Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996). Life with 6000 genes. *Science.* 274, 546–567.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science.* 278, 609–614.

Hoersch, I., Leroy, I., Brown, N. P., Andrade, M. A., and Sander, I. (2000). The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem Sci.* 25, 33–35.

Huynen, M. A., Diaz-Lazcoz, Y., and Bork, P. (1997). Differential genome display. *Trends Genet.* 13, 389–390.

Huynen, M. A., and Snel, B. (2000). Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem*. 54, 345–379.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28, 27–30.

Koonin, E. V. (1997). Genome sequences: genome sequence of a model prokaryote. *Curr. Biol*. 7, R656–R659.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y., and Walker, D. R. (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, 619–637.

Kunst, F., Ogasawara, N., Moszer, I., et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature*. 390, 249–256.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*. 285, 751–753.

Medigue, C., Rechenmann, F., Danchin, A., and Viari, A. (1999). Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics.* 15, 2–15.

Natale, D. A., Galperin, M. Y., Tatusov, R. L., and Koonin, E. V. (2000). Using the COG database to improve gene recognition in complete genomes. *Genetica*. 108, 9–17.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S.A.* 96, 2896–2901.

Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Jr, E. S., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28, 123–125.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 96, 4285–4288.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. and Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234.

Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* 95, 5857–5864.

Snel, B., Bork, P., and Huynen, M. (2000). Genome evolution. Gene fusion versus gene fission. *Trends Genet.* 16, 9–11.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*. 278, 631–637.

Taylor, B. L. and Zhulin, I. B. (1999). PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* 63, 479–506.

Tuininga, J. E., Verhees, C. H., van der Oost, J., Kengen, S. W., Stams, A. J., and de Vos, W. M. (1999). Molecular and biochemical characterization of the ADP-dependent phosphofructokinase from the hyperthermophilic archaeon Pyrococcus furiosus. *J. Biol. Chem.* 274, 21023–21028.

Walker, D. R., and Koonin, E. V. (1997). SEALS: a system for easy analysis of lots of sequences. *ISMB*. 5, 333–339.

Wootton, J. C., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*. 266, 554–71.

16

LARGE-SCALE GENOME ANALYSIS

Paul S. Meltzer

Cancer Genetics Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland

INTRODUCTION

The availability of complete or near-complete catalogs of genes for organisms of increasing complexity has created opportunities for studying numerous aspects of gene function at the genomic level. Gene expression, defined by steady-state levels of cellular mRNA, has emerged as the first aspect of gene function amenable to genome-scale measurement with readily-available technology. It is now possible to carry out massively parallel analysis of gene expression on tens of thousands of genes from a given sample. For model organisms such as *S. cerevisiae*, total genome expression analysis is now routine (Lashkari et al., 1997; Wodicka et al., 1997). For higher eukaryotes, expression measurements that cover a significant proportion of the genome are currently possible, and complete genome expression analysis appears to be an achievable goal. Moreover, this analysis can be repeated on multiple samples, allowing for the statistical analysis of the behavior of genes across a large number of conditions. The massive quantities of data generated by this type of analysis have created new bioinformatics challenges, but these obstacles are well worth overcoming as this type of information has already begun to provide new insights into genome function. So far, this promise has been best realized in *S. cerevisiae*, in which whole-genome measurements have been used to examine fundamental processes such as the cell cycle and the roles of specific transcription factors (DeRisi et al., 1997; Cho et al., 1998; Spellman et al., 1998; Holstege et al., 1998; Chu et al., 1998; Roberts et al., 2000). Although significantly more difficult, similar prob-

lems are now being productively approached in mammalian systems (DeRisi et al., 1996; Amundson et al., 1999; Khan et al., 1999; Iyer et al., 1999; Feng et al., 2000). In addition to their impact on fundamental questions in biology, these technologies are being used to probe the complexity of human diseases, particularly cancer (Khan et al., 1998; Golub et al., 1999; Alon et al., 1999; Alizadeh et al., 2000; Ross et al., 2000; Scherf et al., 2000; Bittner et al., 2000).

Access to large numbers of measurements allows for the statistical analysis of gene expression across multiple samples. In the broadest sense, this opens the possibility of identifying patterns of coregulation among genes, which, in turn, reflects underlying regulatory mechanisms and functional interrelationships. Because mammalian genomes are mainly populated by genes of unknown function, it is theoretically possible to assign anonymous genes to pathways or at least to generate hypotheses as to function by identifying the circumstances that alter their expression. Computational techniques for processing gene expression data are rapidly evolving as many investigators attack the problem from diverse perspectives. The following discussion will present an overview of the technologies for generating and processing expression data, with an emphasis on those techniques and databases that are publicly available.

TECHNOLOGIES FOR LARGE-SCALE GENE EXPRESSION

Measurements

Two broad categories of technology have emerged that can provide large-scale expression data. The first is sequence-based, as exemplified by the serial analysis of gene expression (SAGE) (Velculescu et al., 1995). An alternative hybridization approach is generically termed microarray hybridization (Schena et al., 1995; Lockhart et al., 1996). These two technologies have distinct advantages and disadvantages. In SAGE, a short unique sequence tag is generated from each gene by a PCR-based strategy (Fig. 16.1). Concatemerized tags are sequenced, and the abundance of these tags provides a measurement of the level of gene expression in the starting material. As illustrated below, SAGE tags can be linked to a specific transcript designation in an appropriate database of unique transcripts, such as UniGene. Thus, SAGE is essentially an accelerated technique for cDNA library sequencing. Because it is sequence intensive, SAGE is not well suited to the analysis of large numbers of samples. However, because SAGE does not require a priori knowledge of the pattern of gene expression in a given mRNA source, the same biochemical and bioinformatics procedure can be applied to any sample, given the availability of the appropriate reference database of SAGE tags.

By contrast, in the microarray strategy, labeled cDNAs (the ''target'') are hybridized to an array of DNA elements (the ''probes'') affixed to a solid support (Fig. 16.2). The array elements can either be synthetic oligonucleotides or larger DNA fragments. High densities are achievable and enable the measurement of over 10,000 genes with either type of microarray. In contrast to SAGE, microarray analysis can only measure the expression of genes that correspond to the sequences included in the array fabrication process. Therefore, complete genome expression analysis by microarrays requires the construction of complete microarrays. This difficulty is counterbalanced by the ease of individual experiments that enable the analysis of

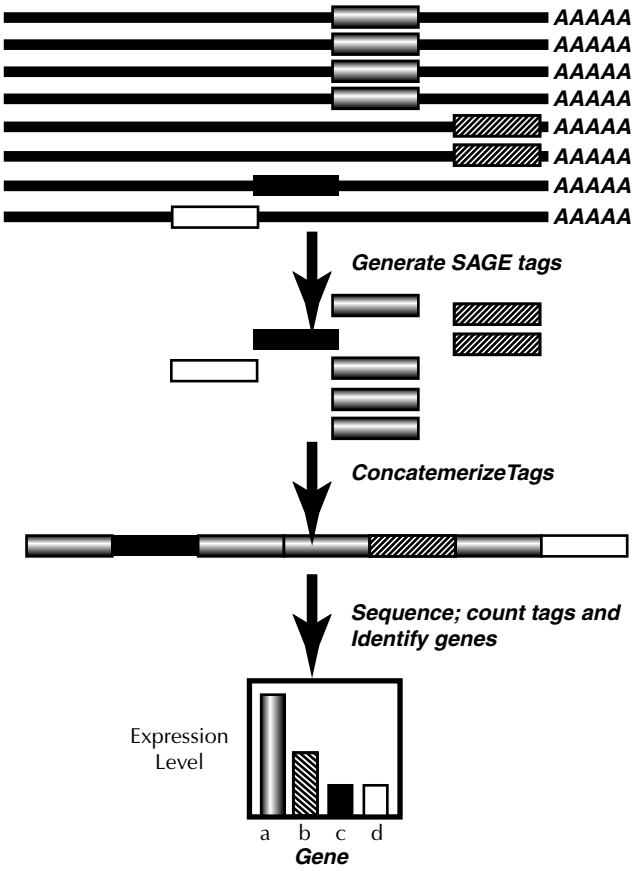


Figure 16.1. Serial Analysis of Gene Expression (SAGE) depends on the generation of a tag from the 3′ end of an mRNA. Tags are concatemerized and sequenced. These data are compared with a database of tags linked to individual transcripts to generate the frequency of each tag in the library, a measure of the expression level for that gene. (See color plate.)

numerous samples. Microarray expression databases are being developed that contain information derived from hundreds or even thousands of samples.

Informatics Aspects of Microarray Production

For organisms with completely sequenced and well-annotated genomes such as *S. cerevisiae*, the production of arrays is relatively straightforward. By selecting a primer pair that amplifies each ORF from genomic DNA, a set of PCR products encompassing the complete genome can be produced. Primer pairs for this purpose are commercially available. Likewise, a series of oligonucleotide array elements can be selected from each ORF. More complex genomes, which may not yet be fully annotated, are more problematic. The consensus at this juncture is to utilize a catalog of genes such as UniGene, generated by reduction of EST sequencing data to unique, clustered transcripts. Clustered ESTs can then be used to predict oligonucleotide sequences for array fabrication or to select a representative EST clone for each

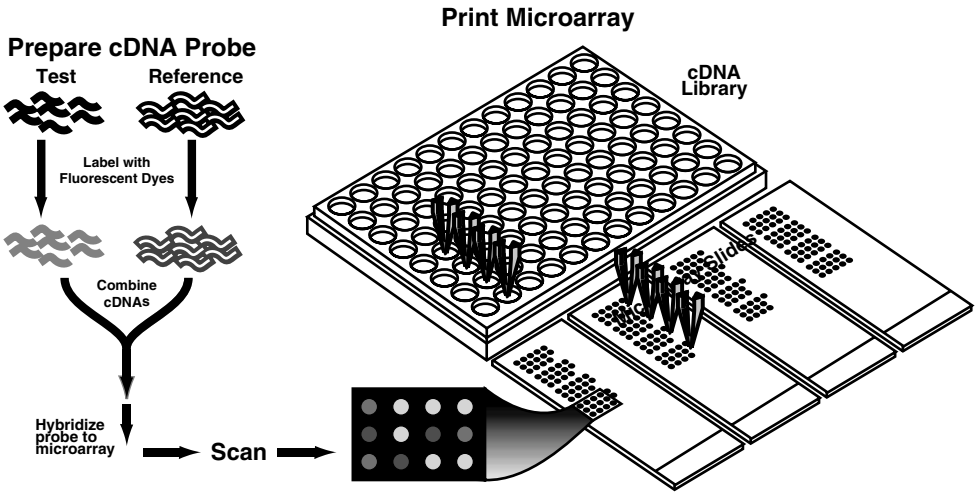


Figure 16.2. The process of microarray hybridization using printed DNA probes. A robotic printer deposits DNA in a regular array on a series of glass slides. After they are processed, the slides are hybridized to a mixture of two cDNA pools derived from test and reference samples that have been labeled with spectrally distinct fluorochromes. After stringency washes, the microarray is scanned in a laser-scanning device, and the image is processed to generate numerical data. (See color plate.)

cluster; in turn, this is then used to generate a cDNA fragment by PCR for deposition on a microarray. This approach is intrinsically limited by the quality and completeness of the EST database, as well as the intrinsic instability of EST cluster designations that evolve along with the EST sequencing projects. A further practical difficulty unique to the cDNA microarray strategy is brought about by the requirement of this system for a physical clone. Recovery of low-redundancy clones may prove difficult due to problems intrinsic to the handling of arrayed libraries. Ideally, EST-based approaches will be supplemented (or even supplanted) by gene catalogs based on genomic sequence that will ultimately provide stable databases for array construction.

In order for the information generated by expression analysis to be useful, it is necessary for each EST to be linked to as much biological information as possible. The accessibility and quality of this information depend on the organism. Because the interpretation of results often relies on the expertise of an individual investigator, issues surrounding gene nomenclature and annotation are significant. For example, an investigator may be familiar with a gene by its common literature alias, but it is cataloged under a more cryptic ''official'' name. ESTs from known genes may fail to be annotated as genes altogether if the available sequence is of insufficient quality to pass the filters required for cluster assembly with known mRNAs. Fortunately, mammalian gene catalogs such as NCBI's LocusLink are being developed that attempt to address these problems.

What is Actually Measured?

A detailed review of array technologies is beyond the scope of this chapter, but the user of expression data must have an understanding of the types of measurement

reported by various expression platforms and incorporated in databases. To understand the options available to mine information in expression data, it is essential to understand the measurements generated by the various technologies presently available. Ideally, an expression measurement would be converted into copies of a given mRNA per cell. Unfortunately, no technology measures this value directly. The variety of expression measurement platforms in use creates problems in the cross-comparison of data generated by technologies and in some cases even within a given technology. For example, SAGE measures the abundance of a particular tag. The validity of this number will depend on the number of tags sequenced. A sample of 1,000 tags will be much less reliable than a sample of 20,000 tags. Hybridization-based systems measure the abundance of cDNAs in a synthetic population of nucleic acids, which is a representation of the mRNA composition of the cell. Variations in biochemical manipulations, hybridization efficiency across array elements, and cross-hybridization may distort the accuracy of measurements. This calls for caution when comparing data generated by different techniques. For example, most synthetic oligonucleotide array assays are based on hybridization of a representation synthesized by amplification from a cDNA template that has been engineered to contain a phage promoter. In contrast, most fluorescent cDNA microarray assays use direct incorporation of tagged nucleotides in the cDNA prepared by reverse transcription of mRNA from the source of interest.

There are few data that directly compare the results of these three assays on the same samples. Moreover, abundance measurements from oligonucleotide array assays are reported as expression levels on a continuous scale. In contrast, printed fluorescent cDNA arrays require the use of a two-color system incorporating a reference mRNA to compensate for variations in the performance of individual array elements and array slides. Although the intensities of each channel are measured and reported, the most robust measurement is expressed as the normalized ratio of intensities for each gene. Radioactive hybridization to cDNA microarrays printed on nylon membranes presents additional problems because only a single channel is measured in a given hybridization, and normalization must rely on cross-comparison of experiments. In addition, expression scales vary in an individual way from linearity and have distinct thresholds and saturation levels depending on the technology and instrumentation utilized. In principle, measurements from these various systems could be converted to a common format, but as yet there are no standards for such a conversion. To be certain that computational techniques are appropriately applied, users of array expression databases and experimentalists venturing into microarray research need to be aware of the characteristics of the data generated by the particular experimental platform in use.

Aspects of the primary analysis of array data are illustrated for printed cDNA microarrays in the following example. This methodology achieves accurate ratio measurements of the relative abundance of each mRNA in the test and reference sources by combining the test cDNA pool with a reference cDNA labeled with a spectrally distinct fluorophore (Fig. 16.2). This is accomplished by obtaining images of the hybridized microarray with a two-channel laser-scanning device for analysis with software (such as DeArray, ScanAlyze, CrazyQuant, or proprietary software bundled with scanner instruments) that measures the signal intensity over background, normalizes the two channels, calculates ratios, generates overall array statistics, and outputs a pseudocolor image and data spreadsheet. This process is illustrated here using DeArray, a module of Array Suite. Examining the scatter plot of

one channel versus the other (Fig. 16.3) is extremely useful. In most hybridizations, the majority of genes will show minimal differences, so the scatter plot would then be centered on the diagonal line. Deviation from this behavior, as illustrated in the example at lower intensities, indicates the signal intensity level at which the data should be filtered. In other cases, the scatter plot might simply tend to widen as intensities drop, again indicating that confidence in ratio measurements declines with intensity. A similar scatter plot constructed from single-channel data from two one-

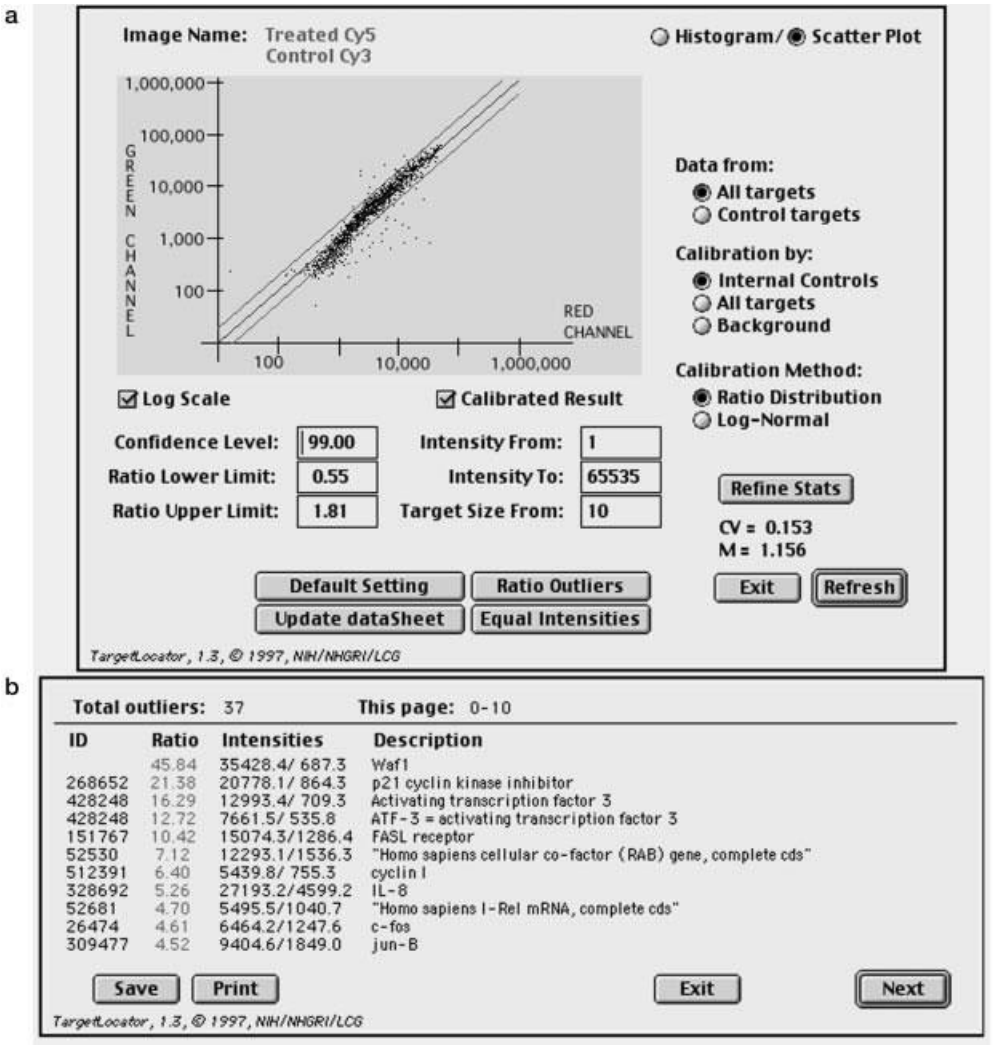


Figure 16.3. (a) Scatter plot illustrating the fluorescent intensity measurements in two channels plotted against each other. Note that most measurements fall on the diagonal. Deviation from this pattern at low intensity indicates that the data are skewed due to nonspecific fluorescence in one channel. For analysis, the data should be filtered to exclude these points. (b) After data are filtered for intensity and a minimum spot size (to exclude spurious measurements), an outlier list can be displayed. A spreadsheet containing the data from the entire microarray can also be saved.

color or radioactive hybridizations is a useful adjunct to the analysis of data generated from these platforms. The analysis of a single microarray experiment is straightforward but not usually very illuminating. The real power of high-throughput gene expression technologies emerges when multiple experiments are subjected to comparison and statistical analysis. To accomplish this goal, expression data must be stored in an appropriate database.

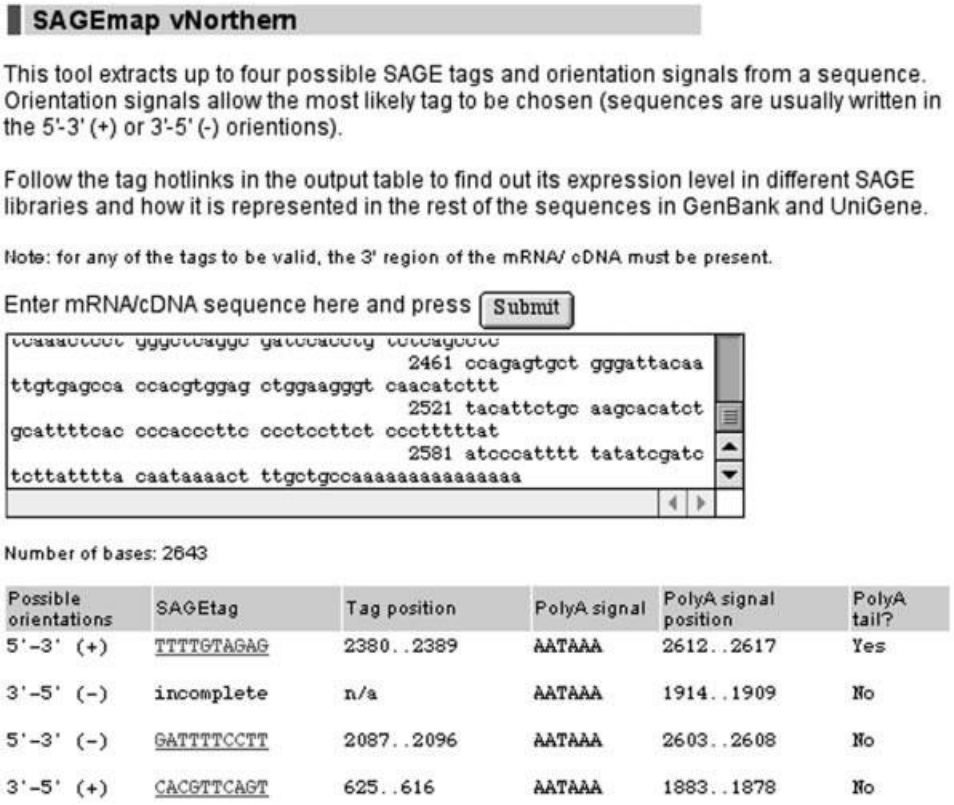
COMPUTATIONAL TOOLS FOR EXPRESSION ANALYSIS

Public Databases

At the present time, the only centralized, publicly maintained repositories of high-throughput gene expression data, aside from EST sequencing databases, contain SAGE data. A single, central repository for all expression data comparable to GenBank is presently not available. This situation may change if standards for the uniform reporting of expression data are adopted. Many useful databases containing the results of various studies are maintained by individual laboratories listed at the end of this chapter.

NCBI maintains a SAGE database called SAGEmap, currently containing data from 69 libraries (Lal et al., 1999). These data can be searched either for data on individual genes or for lists of genes differentially expressed between libraries or pools of libraries. Given a cDNA sequence (which must include the 3′ end), the user can search for its representation in a number of SAGE libraries. SAGE library data are downloadable, and a submission tool for SAGE laboratories is also available. To search the SAGEmap database for data on a given gene using the ''virtual Northern tool,'' the target cDNA sequence is submitted and possible tags are returned (Fig. 16.4). As in the example, one gene may contain multiple tags. Linked to this tag is a display indicating the relative abundance of this tag in a SAGE library with a proportional gray scale image providing a ''virtual Northern'' (Fig. 16.5). In addition, the data are linked to the UniGene cluster(s) corresponding to this tag. The UniGene designation provides a link to an abundance table of all tags for this particular UniGene entry (Fig. 16.6).

Instead of searching for individual genes, SAGEmap can be queried using x-Profiler, a tool that carries out a ''virtual subtraction'' to develop lists of tags that are present in one library or group of libraries at a differing frequency from a reference set. After the user selects the libraries to compare (Fig. 16.7a) and the ratio cutoff between the libraries, xProfiler returns a list (Fig. 16.7b) of differentially expressed genes in the two sources that can be downloaded for further analysis. It should be noted that SAGE data are subject to error, primarily related to sequencing error in the SAGE libraries and in the sequence databases that are used to define gene clusters. The impact of this error is inversely proportional to the number of times a given tag is counted. Tags that occur only once in the library, therefore, represent the least reliable data. xProfiler can be applied in essentially the same fashion to provide a virtual subtraction of EST sequence data compiled as part of the Cancer Genome Anatomy Project (CGAP). Another tool available for screening CGAP and dbEST library data, digital differential display (DDD), provides a gray-scale output indicating the relative abundance of statistically significant differentially expressed genes.



SAGEmap vNorthern

This tool extracts up to four possible SAGE tags and orientation signals from a sequence. Orientation signals allow the most likely tag to be chosen (sequences are usually written in the 5'-3' (+) or 3'-5' (-) orientions).

Follow the tag hotlinks in the output table to find out its expression level in different SAGE libraries and how it is represented in the rest of the sequences in GenBank and UniGene.

Note: for any of the tags to be valid, the 3' region of the mRNA/ cDNA must be present.

Enter mRNA/cDNA sequence here and press Submit

```
                                    2461 ccagagtgct gggattacaa
ttgtgagcca ccacgtggag ctggaagggt caacatcttt
                                    2521 tacattctgc aagcacatct
gcattttcac cccacccttc ccctccttct ccctttttat
                                    2581 atcccatttt tatatcgatc
tcttatttta caataaaact ttgctgccaaaaaaaaaaaaaaa
```

Number of bases: 2643

Possible orientations	SAGEtag	Tag position	PolyA signal	PolyA signal position	PolyA tail?
5'-3' (+)	TTTTGTAGAG	2380..2389	AATAAA	2612..2617	Yes
3'-5' (-)	incomplete	n/a	AATAAA	1914..1909	No
5'-3' (-)	GATTTTCCTT	2087..2096	AATAAA	2603..2608	No
3'-5' (+)	CACGTTCAGT	625..616	AATAAA	1883..1878	No

Figure 16.4. Searching SAGEmap for tags representing a given gene ("virtual Northern"). The example illustrates the cDNA sequence of human p53 (*TP53*). Potential tags and their positions in the p53 mRNA are returned.

Most public microarray data are accessible only through individual laboratory Web sites, many of which are listed at the end of this chapter. It can be anticipated that unified public expression databases will be developed once issues as to data format and cross-platform comparison are resolved. One of the preliminary efforts currently under development and aimed at the public use and dissemination of gene expression data is the NCBI Gene Expression Omnibus (GEO). This database is intended to house different types of expression data, including oligonucleotide and cDNA microarray data, hybridization filter data, and SAGE data. Although this platform will undoubtedly evolve as more and more gene expression information becomes available, GEO is currently envisioned as having four primary entities:

1. Submitter, which contains contact and login information on the submitter,
2. Platform, which contains information on the physical reagents used in the actual experiment,
3. Sample, which deals with the mRNA samples in question and the data generated from the actual experiment, and

SAGE Tag to Gene Mapping

SAGEtag (10 bases): TTTTGTAGAG

Reliable UniGene clusters matched to this tag:

Hs.1846 : **tumor protein p53 (Li-Fraumeni syndrome)**

SAGE library data for this tag:

Library name	*Tags per million*		*Tag counts*	*Total tags*
SAGE Duke BB542 normal cerebellum	24		1	41450
SAGE Chen LNCaP	31		2	62681
SAGE Chen LNCaP no-DHT	15		1	65206
SAGE Chen Normal Pr	29		2	66687
SAGE Duke GBM H1110	42		3	71138
SAGE SW837	32		2	61290
SAGE RKO	19		1	52423
SAGE CPDR LNCaP-C	95		4	41848
SAGE CPDR LNCaP-T	45		2	44370
SAGE pooled GBM	15		1	63078
SAGE BB542 whitematter	20		2	96344
SAGE Panc 91-16113	58		2	34159
SAGE OVCA432-2	345		1	2894
SAGE OV1063-3	50		2	39466
SAGE Duke H341	44		2	44983
SAGE OVP-5	359		4	11118
SAGE OVT-6	69		3	43074
SAGE OVT-7	18		1	55476
SAGE SKBR3	119		1	8379
SAGE Duke HMVEC	37		2	53089
SAGE Duke HMVEC+VEGF	17		1	58499
SAGE DCIS	24		1	41540
SAGE Duke 757	50		1	19674
SAGE OVT-8	29		1	34096
SAGE A2780-9	88		2	22551
SAGE ML10-10	122		7	57326
SAGE Duke H247 normal	16		1	61096
SAGE A+	47		1	21231
SAGE IOSE29-11	122		6	48876

Number of SAGE libraries: 69
Total tags in all SAGE libraries: 2877861

Figure 16.5. Following the link from a given tag in the virtual Northern results (Fig. 16.4) provides a display of the abundance of the tag in the database with a gray scale "virtual Northern" to facilitate scanning the long list.

SAGE Gene to Tag Mapping

UniGene cluster: Hs. 1846 Submit

Hs.1846 : tumor protein p53 (Li-Fraumeni syndrome)

SAGE library data and reliable tag summary:

Reliable tags found in SAGE libraries:

GGGTCTAGAA	*Library name*	*Tags per million*	*Tag counts*	*Total tags*
	SAGE Duke H1020	18	1	53554
	SAGE 293-CTRL	44	2	44667
	SAGE Chen LNCaP	47	3	62681
	SAGE Chen Normal Pr	44	3	66687
	SAGE Chen Tumor Pr	14	1	69202
	SAGE Duke H392	34	2	58099
	SAGE Duke GBM H1110	14	1	71138
	SAGE NHA(5th)	18	1	53219
	SAGE Tu102	17	1	58190
	SAGE Tu98	20	1	49527
	SAGE Duke H341	22	1	44983
	SAGE ES2-1	31	1	31763
	SAGE LNCaP	43	1	22935
	SAGE Duke HMVEC+VEGF	17	1	58499
	SAGE A2780-9	44	1	22551
	SAGE ML10-10	17	1	57326
	SAGE Br N	26	1	38274
	SAGE IOSE29-11	40	2	48876
TGCATTTTCA	*Library name*	*Tags per million*	*Tag counts*	*Total tags*
	SAGE 293-CTRL	22	1	44667
	SAGE Duke mhh-1	20	1	48959
	SAGE SciencePark MCF7 control 3h	168	1	5924
TTCAAGACAG	*Library name*	*Tags per million*	*Tag counts*	*Total tags*
	SAGE IOSE29-11	20	1	48876
TTTTGTAGAG	*Library name*	*Tags per million*	*Tag counts*	*Total tags*
	SAGE Duke BB542 normal cerebellum	24	1	41450
	SAGE Chen LNCaP	31	2	62681
	SAGE Chen LNCaP no-DHT	15	1	65206

Figure 16.6. The complete list of tags from the virtual Northern results (Figs. 16.4 and 16.5) can also be used to query the database, finding the occurrence of their tags in all SAGE libraries.

4. Series, which houses information on collections of samples and the relationship between the samples. The series entity will also contain the results of any data clustering.

Even when public databases for microarray expression data are established, investigators who are setting up microarray laboratories soon learn that data storage is an essential requirement for their research. Sources for freely available software for expression databases are listed at the end of this chapter. One example of such a

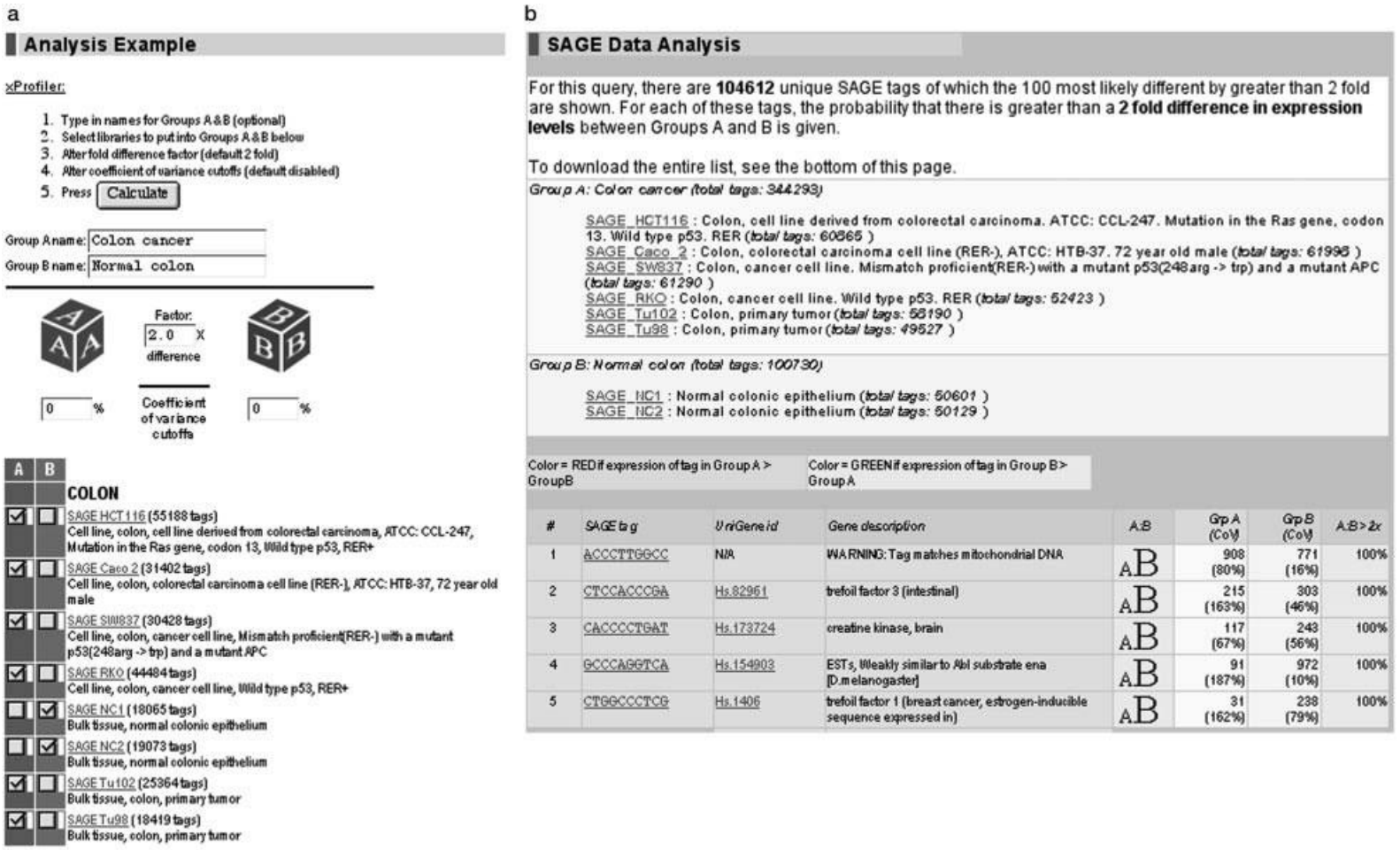


Figure 16.7. (a) The xProfiler tool allows the user to perform an electronic subtraction between sets of SAGE libraries by selecting libraries in this window. In this example, normal colonic epithelium is subtracted from colon cancer. (b) The results of the electronic subtraction are displayed in tabular form.

database is ArrayDB (Masiello et al., 1999). ArrayDB, which requires either a Sybase or Oracle client, was designed for printed microarray data and allows the storage of experimental data and simple data queries. ArrayDB is designed to store a database of clones used for array fabrication and the data output from individual hybridization experiments. Links to the UniGene database are maintained. Data can be downloaded or viewed through a Web-based Java applet. On entry to the database, one selects the experiment to view from a pull-down menu and, on loading, a histogram of ratios appears (Fig. 16.8). This window contains several selectable options for viewing portions of the data. Importantly, the data can be filtered according to intensity and spot size to remove insignificant measurements from subsequent analyses. On submitting a query, a new window opens providing an image of the array and a table of genes that match the query results (Fig. 16.9). The image is useful because it allows confirmation that a given value is not the consequence of hybridization artifact such as a scratch or dye precipitate. The table contains the relevant intensity, spot size, and ratio data for each gene meeting the search criteria. Clicking on the Clone ID field in this view opens a window containing data on the clone printed at that point on the array with links to relevant databases including UniGene, OMIM, GenBank, and GeneCards, which are useful for acquiring functional information about a given gene. Additional features of ArrayDB include the ability to see a list

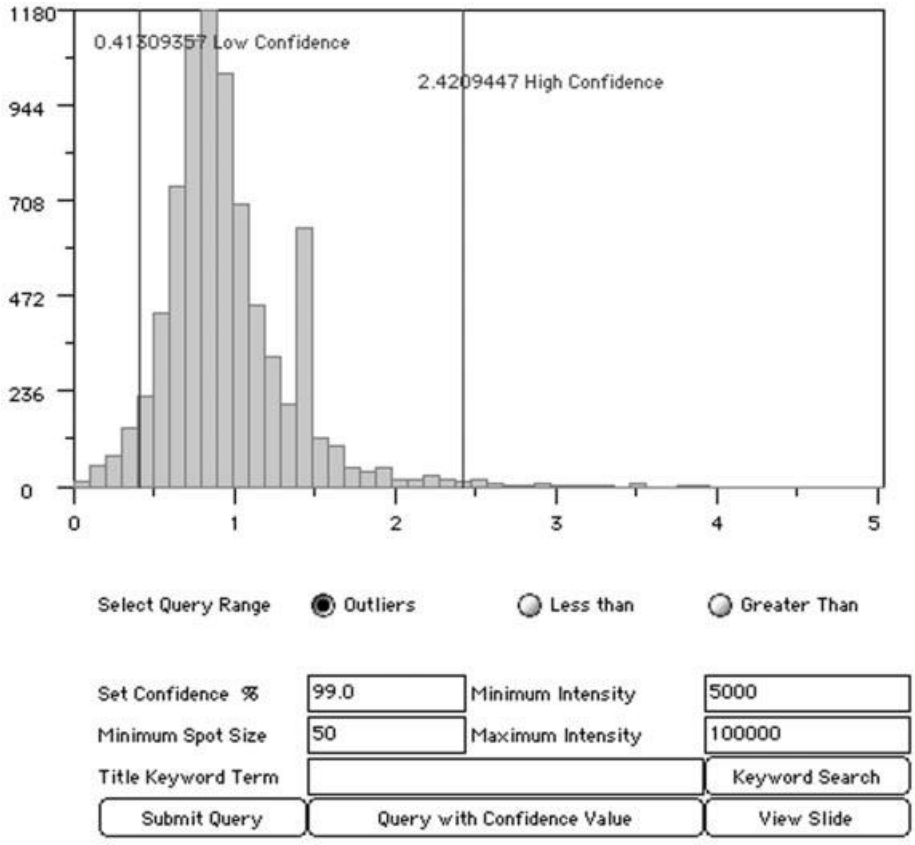


Figure 16.8. Ratio histogram of a microarray hybridization retrieved from ArrayDB.

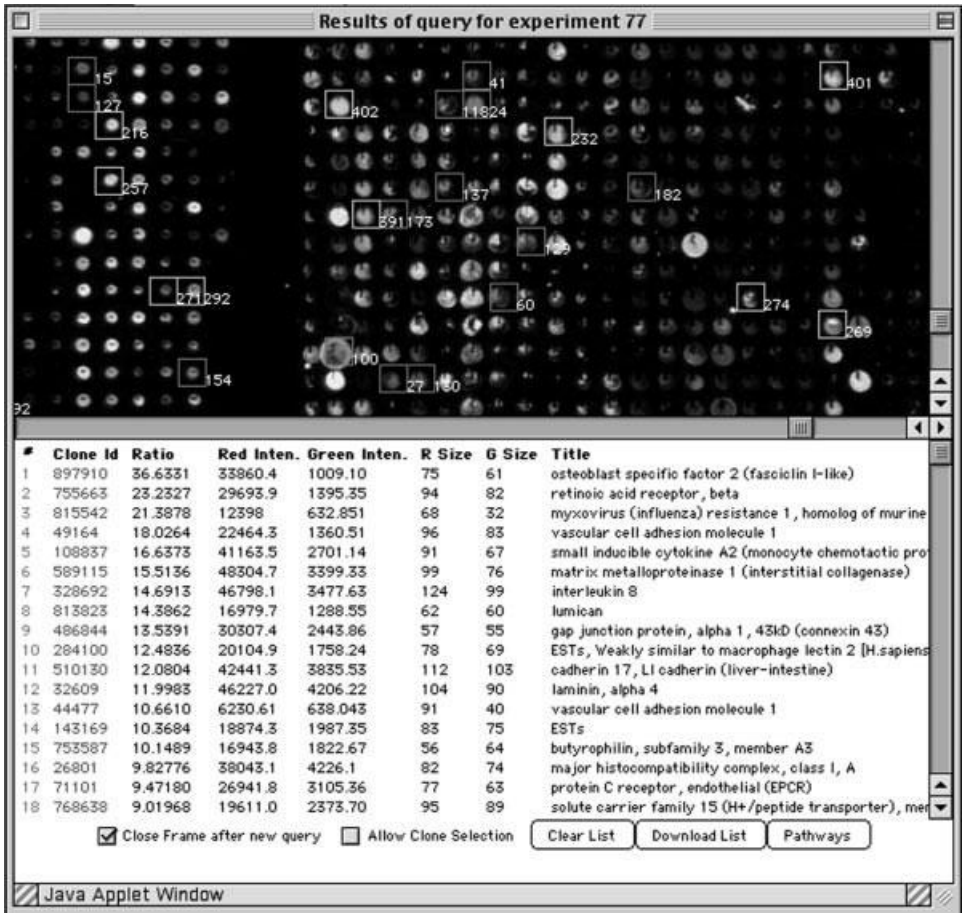


Figure 16.9. In ArrayDB, query for outliers returns an image of the microarray, with outlying genes highlighted in the image and listed below the image, along with intensity data and clone identifiers. (See color plate.)

of biological pathways that gene products of interest are involved in; the positions of these gene products within known biological pathways are also shown graphically using image maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG). Investigators involved in microarray research soon learn the critical importance of rapid access to descriptive information on gene function. Most of the genes of interest in individual experiments are unfamiliar, and an efficient mechanism for obtaining a synopsis of gene function is essential to the interpretation of microarray data.

Central storage of microarray data with investigator access via individual work stations is clearly the preferred mode for information management for centers generating significant quantities of microarray data and ultimately will be required for public distribution of expression data. However, it is important to note that projects of significant size can be accommodated in low-cost databases maintained on desktop computers. As an example of this approach, FileMaker Pro templates for cDNA microarray data can be downloaded from the NHGRI microarray site. Projects of up to 50 experiments can easily be accommodated. Although, as discussed below, the

statistical analysis of expression data with specialized tools is important, the value of simple relational databases should not be underestimated. Even in a database such as FileMaker Pro, it is possible to define patterns of genes expressed consistently across experiments and possible to generate useful graphs displaying numerical data as color blocks (Fig. 16.10). This type of graphic display is much easier to scan by eye than a large table of numbers. Reanalyzing a data set using various filters is greatly facilitated by placing the entire set of experiments in a searchable database. Various settings can be tested until the output is optimized.

Initial processing of data requires the application of significance filters to the raw data set so that only meaningful measurements enter into downstream analysis. This requires the application of a sensitivity threshold to remove genes that have not been measured accurately. The importance of this step was illustrated earlier in this chapter. Individual laboratories will need to establish thresholds applicable to their own data, and investigators should exercise caution when using publicly accessible databases to be certain whether the data has been filtered for a detection threshold. Of course, there is no fixed cutoff that must be applied to all expression data, but, in general, as the threshold is lowered toward the background noise of an assay, the quality of the data will diminish. To maximize the yield of information from an experiment and to simplify data displays by removing uninformative genes, in most cases it is also useful to apply a filter that removes genes that, although measured accurately, do not fluctuate across a series of experiments.

Once an appropriately filtered data set has been extracted, the data are ready for analysis. In some instances, simple searches may be sufficient to generate lists of genes that are expressed under a given condition. For example, from a set of experiments in which cells have been treated with a drug or transfected with a gene, simple searching and sorting of the filtered data will yield a list of genes induced or

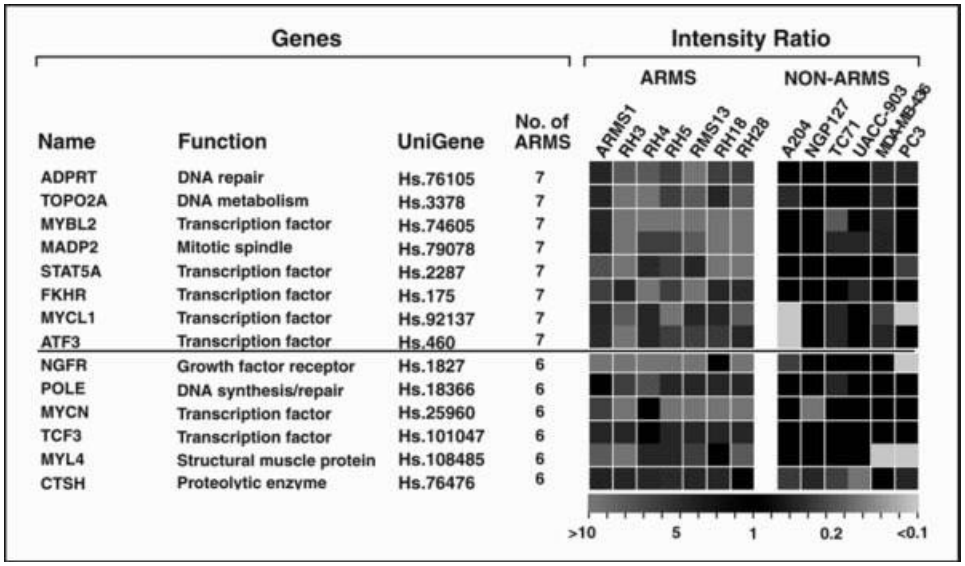


Figure 16.10. Display of microarray results retrieved from FileMaker Pro. This example illustrates the results of a query for genes upregulated in a series of cancer cell lines with ratios coded by a red to green color map (Khan et al., 1998). (See color plate.)

repressed by the treatment. However, the limitations of simple searches are quickly reached with larger data sets involving multiple samples or conditions and computational techniques for data organization are essential.

Cluster Analysis

The goal of cluster analysis is to reveal underlying patterns in data sets that contain hundreds of thousands of measurements and to present this data in a user-friendly manner. Ideally, patterns of similarities and difference among samples are identified and genes are coregulated in distinct patterns. A number of tools for clustering have been developed, which are freely available. In general, these are based on conventional statistical techniques widely used in other contexts and are largely dependent on linear correlation analysis. It is certain that future research efforts will be directed at the development of increasingly sophisticated tools designed for mining expression data. By applying statistical analysis, definition of subsets in apparently homogeneous tissue samples, development of classifiers for various disease entities, and identification of groups of coregulated genes may be possible. These possibilities are especially intriguing because they may provide a way to assign the large number of anonymous genes in higher eukaryotes to functional groups.

HIERARCHICAL CLUSTERING

Agglomerative hierarchical clustering has established itself as the most frequently applied technique for processing array data for inspection (Weinstein et al., 1997; Khan et al., 1998; Eisen et al., 1998). All pairwise comparisons of expression levels are made between experiments (Fig. 16.11). The resulting matrix of scatter plots can be reduced to a matrix of Pearson correlation coefficients. This is readily displayed in two dimensions as a hierarchical dendrogram (Fig. 16.12). Both genes and samples can be clustered in this fashion. By color coding expression levels, a large numerical table can be replaced by a much more compact and easily inspected color plot

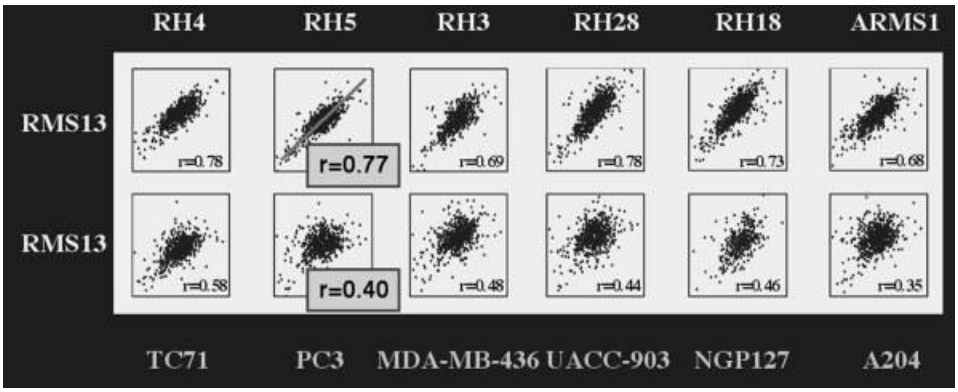


Figure 16.11. Portion of a scatter plot matrix illustrating pairwise plots of log ratios across a series of microarray experiments. Each scatter plot is used to calculate a Pearson correlation coefficient between the two sets of measurements (Khan et al., 1998).

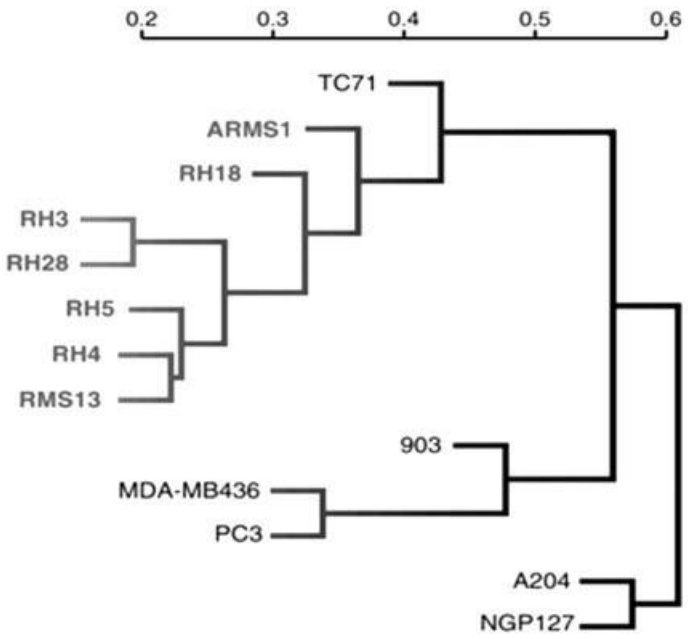


Figure 16.12. Hierarchical clustering dendrogram plotted from the Pearson correlation coefficients calculated across a series of experiments (Khan et al., 1998).

(Weinstein et al., 1997; Eisen et al., 1998). Inspection of the dendrogram and color plot will reveal samples with closely related patterns of gene expression as well as clusters of genes with similar expression pattern. Software for this type of display is freely available. This approach has been productively applied to yeast and human expression data. It is important to note that there is an arbitrary character to the manner in which a dendrogram is drawn. Clusters can be rotated about the point of bifurcation affecting the apparent proximity of the edges of a cluster with adjacent clusters. The important information is contained in the cluster contents and their similarity.

An alternative approach for displaying the same information is multidimensional scaling (MDS) (Fig. 16.13). MDS software is available in standard statistical software such as MATLAB. The distance measure for plotting individual samples is based on $1 - r$ where r is the Pearson correlation coefficient. Thus samples that are closely related will plot closely together, and widely differing samples plot farther apart. Although MDS does not allow simultaneous display of gene and sample clusters, it has the virtue of retaining a graphical display that places each sample (or gene) plotted in relation to every other. Ideally, an MDS plot, which is a reduction of multidimensional data to a three-dimensional display, should be viewed on a computer screen that allows rotation of the data so that the viewer is not deceived by any single two-dimensional projection.

Several other statistical tools based on linear and nonlinear methods have been used to analyze array data, including self-organizing maps, k-means clustering, and principal component analysis (Tamayo et al., 1999; Ben-Dor et al., 1999). Each of these tools has merit, and software for some of these is publicly available. However,

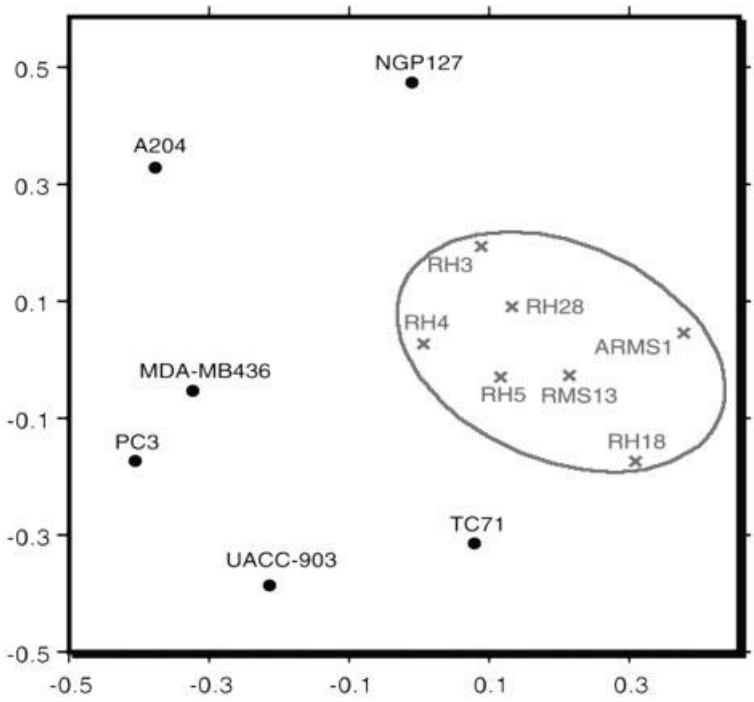


Figure 16.13. Multidimensional scaling plot of the same data represented in Figure 16.12. Note that the samples that fall closely together in the dendrogram also plot near to one another in the MDS plot. Ideally, this plot should be viewed in three dimensions on a computer screen so that it can be rotated to allow better inspection of the distances between the data points.

the precise selection of statistical methods for the analysis of array data is still the subject of active investigation, and the limits of these techniques are not well defined (Bittner et al., 1999). No single method can be recommended at this time to the exclusion of others, and most investigators will base their choice on the availability of software tools and Web-based resources. The range and quality of these tools will certainly improve rapidly in the coming years.

PROSPECTS FOR THE FUTURE

Existing technologies for high-throughput gene expression analysis and the informatics approaches to analysis of data arising from these methods are already producing an extremely interesting and novel view of genome function. However, it is apparent that there are limitations to current approaches that present the opportunity for significant improvements. At the level of technology, it can be anticipated that microarrays will move progressively closer to whole genome analysis with improved sensitivity and reduced sample requirements. At the level of informatics, several areas of progress can be foreseen. Tools for providing precomputed informative summaries of function for named genes are likely to be developed. Methods of linking gene

expression data to genomic sequence would be of great value in dissecting networks of coregulated genes by identifying regulatory motifs shared by clustered genes. Most far-reaching would be methods of defining gene networks in terms of codetermination through the development of computational tools that predict the expression of a gene based on the expression of other genes (Seungchan et al., 2000). It is hoped that the recognition of gene networks will facilitate the assignment of function to anonymous genes and the definition of pathways. Ultimately, models may be developed that accurately reflect the mesh of self-compensating regulatory networks that maintain ordered genome function.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 16

NCBI Databases and Tools	
DDD	*http://www.ncbi.nlm.nih.gov/CGAP/info/ddd.cgi*
NCBI Gene Expression Omnibus	*http://www.ncbi.nlm.nih.gov/geo*
SAGEmap	*http://www.ncbi.nlm.nih.gov/SAGE/*
xProfiler	*http://www.ncbi.nlm.nih.gov/CGAP/hTGI/xprof/cgapxpsetup.cgi*
Software	
ArrayDB	*http://genome.nhgri.nih.gov/arraydb/*
Cluster, TreeView, and ScanAlyze	*http://rana.Stanford.EDU/software/*
CrazyQuant	*http://chroma.mbt.washington.edu/mod_www/tools/index.html*
GeneCluster	*http://www.genome.wi.mit.edu/MPR/*
GeneX	*http://www.ncgr.org/research/genex/*
P-SCAN	*http://abs.cit.nih.gov/main/pscan.html*
ScanAlyze; AMAD	*http://www.microarrays.org/*

An up-to-date list of the links to laboratories involved in microarray technology development can be found at the NHGRI Web site (*http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/*).

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Staudt, L. M., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96, 6745–6750.

Amundson, S. A., Bittner, M., Chen, Y., Trent, J., Meltzer, P., and Fornace, A. J., Jr. (1999). Fluorescent cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses. *Oncogene* 18, 3666–3672.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.* 6, 281–97.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, C., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. M. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406, 536–540.

Bittner, M., Meltzer, P., and Trent, J. (1999). Data analysis and integration: of steps and arrows. *Nat. Genet.* 22, 213–215.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 2, 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast [published erratum appears in Science 1998 Nov 20; 282(5393):1421]. *Science* 282, 699–705.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14, 457–460.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.

Feng, X., Jiang, Y., Meltzer, P., and Yen, P. M. (2000). Thyroid hormone regulation of hepatic genes in vivo detected by complementary DNA microarray. *Mol. Endocrinol.* 14, 947–955.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87.

Khan, J., Bittner, M. L., Saal, L. H., Teichmann, U., Azorsa, D. O., Gooden, G. C., Pavan, W. J., Trent, J. M., and Meltzer, P. S. (1999). cDNA microarrays detect activation of a myogenic transcription program by the PAX3-FKHR fusion oncogene. *Proc. Natl. Acad. Sci. USA* 96, 13264–13269.

Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S. B., Pohida, T., Smith, P. D., Jiang, Y., Gooden, G. C., Trent, J. M., and Meltzer, P. S. (1998). Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* 58, 5009–5013.

Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. (1999). A public database for gene expression in human cancers. *Cancer Res.* 59, 5403–5407.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94, 13057–13062.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680.

Masiello, A. J., Chen, Y., Bittner, M. L., Meltzer, P. S., Trent, J. M., and Baxevanis, A. D. (1999). ArrayDB 2.0: Software for the Exploration and Analysis of Microarray Gene Expression Data. Cold Spring Harbor Meeting on Genome Sequencing and Biology, Cold Spring Harbor, New York, p. 151.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., and Friend, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24, 227–235.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. (2000). A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244.

Seungchan, K., Dougherty, E. K., Chen, Y., Krishnamoorthy, S., Meltzer, P., Trent, J. M., and Bittner, M. (2000). *Multivariate measurement of gene expression relationships. Genomics* 67, 201–209.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96, 2907–2912.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270, 484–487.

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Jr., Kohn, K. W., Fojo, T., Bates, S. E., Rubinstein, L. V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. W., Bunow, B., Viswanadhan, V. N., Johnson, G. S., Wittes, R. E., and Paull, K. D. (1997). An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae. Nat. Biotechnol.* 15, 1359–1367.