Functional genomics and proteomics: charting a multidimensional map of the yeast cell☆

Gary D. Bader[1], Adrian Heilbut[2], Brenda Andrews[3], Mike Tyers[3,4], Timothy Hughes[3,5] and Charles Boone[3,5]

[1]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 460, New York, NY 10021, USA
[2]MDS Proteomics, 251 Attwell Drive, Toronto, ON, Canada M9W 7H4
[3]Department of Medical Genetics and Microbiology, University of Toronto, 1 Kings College Circle, Toronto ON, Canada M5S 1A8
[4]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University Avenue, Toronto ON, Canada M5G 1X5
[5]Banting and Best Department of Medical Research, University of Toronto, 112 College St. Toronto ON, Canada M5G 1L6

The challenge of large-scale functional genomics projects is to build a comprehensive map of the cell including genome sequence and gene expression data, information on protein localization, structure, function and expression, post-translational modifications, molecular and genetic interactions and phenotypic descriptions. Some of this broad set of functional genomics data has been already assembled for the budding yeast. Even though molecular cartography of the yeast cell is still far from comprehensive, functional genomics has begun to forge connections between disparate cellular events and to foster numerous hypotheses. Here we review several different genomics and proteomics technologies and describe bioinformatics methods for exploring these data to make new discoveries.

Charting the cell map – that is, how all of the parts of the cell exist, interact and react over space and time – is an enormous challenge for contemporary biology. New experimental strategies combined with complete genomic information and automation technology are allowing biologists to explore cellular function systematically [1,2]. Each large-scale study, from genome sequencing to molecular interaction network mapping, provides knowledge that enables further directed and discovery-based research. This mode of analysis can be likened to mapping based on satellite images, in which a high-altitude view of a geographical region highlights general features that can be surveyed in more detail.

A systematic approach to cell biology first requires an ordered list of parts, such that protein and gene function can be classified in general terms. A more complex goal is to collect, on a large scale, quantitative information such as expression levels of mRNA and protein, rate constants and stoichiometry for biochemical reactions. Such datasets can provide detailed insight into specific cellular functions, for example biological pathways, through rigorous mathematical modeling [3,4], and an integration of this information can enable computational simulation of more general cellular processes, for example cell division [5]. Because cellular processes are often determined by functional modules such as molecular complexes, signaling pathways and whole organelles [6], it is possible to study these modules separately and then integrate them back into a complete system using a systems biology approach [7]. Other approaches that consider stochastic cellular processes [8] are probably also required to understand fully the workings of the cell. To create a meaningful output, the information collected for each approach should be of high quality [9] and must be organized into databases in structured formats that can be interrogated computationally in order to manage, integrate, analyze and visualize all of the data.

The completion of whole genome sequences has greatly accelerated the pace of biological discovery. An illustration of this effect can be seen in the publications on budding yeast, for which the number of papers published per year, describing specific discoveries in many diverse areas, increased enormously between 1992 and 1996 (Fig. 1) when the genome sequence was released [10,11]. We anticipate another substantial jump in discovery rate with the population of large-scale functional databases, such as the Biomolecular Interaction Network Database (BIND) [12], the Database of Interacting Proteins (DIP) [13], the Molecular Interactions Database (MINT) [14], the General Repository for Interaction Datasets (GRID) [15], the MIPS Comprehensive Yeast Genome Database (CYGD) [16] and the Saccharomyces Genome Database (SGD) [17]. These databases are only just starting to be filled and the biological significance of much of the data remains to be validated. For example, although 15 000 of an estimated 30 000 [18] direct physical interactions have been identified, many of these are likely to be false positives [18,19]. In a second example, putative binding sites in the genome for most known and predicted transcription factors have been identified [20–22], but direct regulation has not been demonstrated for most and, furthermore, there is only

☆ This article is the fifth review in our 'Interdisciplinary Biology' series that commenced in the January 2003 issue of *TCB*. *Eds*

Corresponding author: Charles Boone (charlie.boone@utoronto.ca).

0962-8924/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved. doi:10.1016/S0962-8924(03)00127-2

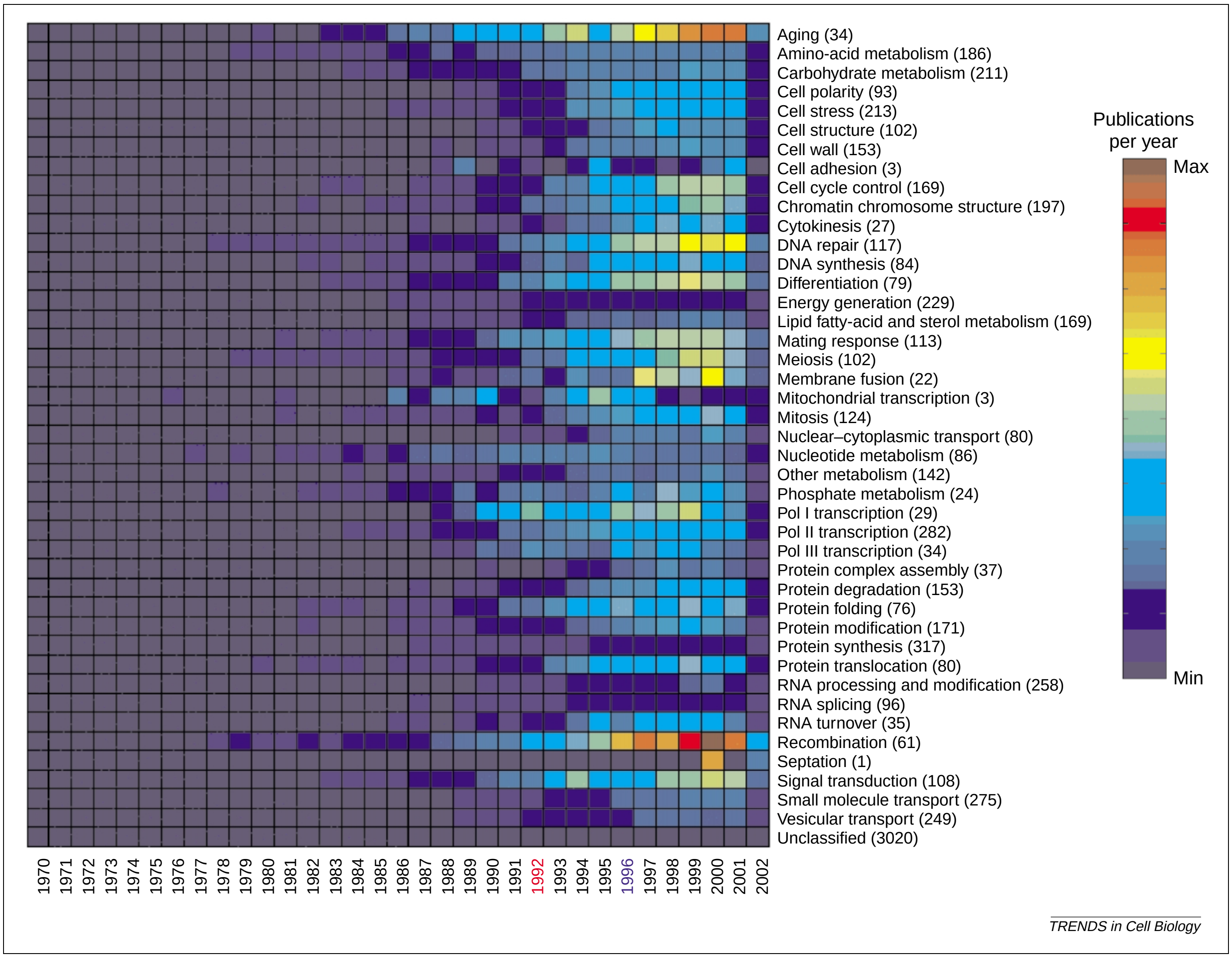


Fig. 1. Publication density by year and by Yeast Proteome Database categories of cellular role. Shown is the increase in the average number of papers per gene per functional category since 1970. Red indicates more papers published and blue indicates less. The number in parentheses after the functional category is the number of genes in each category. Number of publications per gene per year was determined by gene name occurrence (considering all aliases) in a compiled set of 24 000 Medline abstracts listed in the SGD database [17] and in additional Medline abstracts identified by the association of any of the aliases of each yeast gene name together with the strings 'yeast', 'sacch' or 'cerev'. The publications per category are normalized to the number of genes in the category, thus the values shown are normalized units of zero and above and are not the actual number of papers. The first complete yeast chromosome sequence was published in 1992 [10] (red in the *x* axis) and the yeast genome was assembled in 1996 [11] (blue in the *x* axis). It can be seen that a large increase of publications mentioning yeast genes in their abstract occurred in conjunction with the availability of the yeast genome sequence.

minimal overlap between the datasets, probably because of lack of sensitivity [20–22]. Finally, the first systematic analysis of yeast genetic interactions suggests that only a fraction of genetic interactions have been documented so far [23].

The budding yeast is likely to be the first eukaryotic cell to be computationally modeled successfully because of the powerful molecular and genetic methodologies available and the number of systematic large-scale studies currently underway and planned. This modeling might take many decades to complete because of the enormous number of individual reactions and reaction parameters that must be carefully measured for every cell part and among all parts of a complex or pathway [5,24]. Flux balance analyses, which can model whole cells, are easier to construct because they do not require reaction parameter measurements, but they can predict only the limits of normal cellular function and not exact metabolic behaviour [25]. Here we review work completed and in progress to chart the yeast cell, focusing on the elucidation and integration of gene expression patterns and protein–protein and genetic interaction networks in yeast.

Genome sequence

Mapping and sequencing genomes [26,27] are prerequisites for systematic genomics and proteomics. Genes are predicted from the genome, translated to proteins and then functionally annotated on the basis of their similarity to known proteins in databases [28]. Computationally annotating gene function in this manner can also lead to a higher level of understanding; for example, metabolic networks have been partially reconstructed from this type of analysis [29,30]. Unfortunately, the requirement of exon and splice site identification in eukaryotes means that gene prediction is often uncertain and atypical genes can be missed [31]. Consequently, many genes are designated as hypothetical open reading frames (ORFs).

The prediction of genes encoding RNA is generally more difficult, and current identification methods require comparisons of sequenced genomes of organisms that are closely related but have diverged just enough that conserved sequences are differentiated from background [32]. Even the yeast genome, which was assembled in 1996 [11], is still not completely annotated. The complement of yeast genes is undergoing continual refinement as false genes are removed and novel ORFs are added [33]. As each gene and protein is verified as being expressed, the genome becomes more complete. Large-scale application of DNA microarrays to identify expressed exons [34], mass spectrometry to identify expressed proteins [35,36] and polymerase chain reaction (PCR) to identify predicted ORFs [37] can verify genes and their translated proteins in a high-throughput manner.

Phenomics: large-scale gene deletion and RNA-mediated interference

Systematic mutational analysis of every predicted gene offers the potential to assess all genes for a role in a particular biological process using phenotypic analysis. The set of all mutant phenotypes can be defined loosely as the 'phenome' [38]. For yeast, a complete set of deletion mutants has been constructed by PCR-based homologous recombination [39]. This project was carried out by an international consortium of laboratories, which identified about 1000 essential genes and generated roughly 5000 viable haploid gene deletion mutants. The whole set of mutants has been made publicly available, enabling a systematic and comprehensive approach to phenotypic analysis. The power of this approach has been demonstrated by several screens of the set of 5000 viable gene deletion mutants for defects in drug sensitivity [40], cell size [41], cell morphology [42], cell surface function [43], bud site selection [44] and vacuolar protein sorting [45].

'DNA bar codes' – two unique 20-nucleotide oligomers of DNA sequence flanked by common PCR primer sites [39] – are engineered into each deletion cassette and thus unambiguously identify each mutant yeast strain in the collection. Because these bar codes can be detected by hybridization to a bar code DNA microarray, the presence or absence of each deletion strain in a mixed population can be deciphered simply by examining the bar code pattern of a population sample [42]. Pools of diploid strains that are heterozygous for a deletion mutation can be examined – for example, for hypersensitivity to compounds that inhibit growth – in relatively small culture volumes, thereby providing a high-throughput system for linking compounds to their intracellular targets [42]. Application of this analysis to fungal pathogens should facilitate the identification of antifungal drug leads for fungal-specific essential genes [46]. Alternatively, mapping specific phenotypes to genes conserved from yeast to humans might help to identify candidate genes linked to disease. For example, candidate human disease genes associated with mitochondrial defects have been mapped simply by examining the set of 5000 viable deletion mutants for growth defects on a nonfermentable carbon source [47].

In metazoan organisms, RNA-mediated interference (RNAi) offers the potential for systematic phenome mapping by the selective 'knock down' of gene expression. Large-scale analysis of the organismal phenotypes associated with RNAi-based inhibition of *Caenorhabditis elegans* genes has been reported recently [48,49]. Furthermore, the introduction of RNAi constructs into mammalian stem cells, which can be then grown into tissues or adult organisms in which the interfering RNA is expressed in every cell, will vastly accelerate phenotypic screens.

Large-scale screens of mouse RNAi mutants, traditional knockout mutants [50] and chemically mutagenized mutants [51] will enable the measurement of phenotypes in blood and tissue tests, whole-body magnetic resonance imaging, and learning and memory tests, thereby facilitating the elucidation of gene function and the generation of new mouse models of human disease (see TBASE: http://tbase.jax.org/). Finally, the use of microarrays of double-stranded RNA on glass slides for RNAi transfection of many types of cell simultaneously will allow high-throughput phenotypic analysis at a cellular [52] or tissue [53] level. From the perspective of drug discovery, whole chips of cells or grown tissues, each with a different known genetic defect introduced by RNAi, could be used in small-molecule screens.

Transcriptional profiling

Large-scale gene expression analysis with microarrays is a powerful genomics methodology that can be applied to any organism for which the genome has been sequenced or for which extensive cDNA collections have been built [54,55]. As genome sequencing becomes more efficient, the application of highly flexible rapid oligonucleotide synthesis technology such as inkjet [56] and dynamic light-directed [57] synthesis, which provide the ability to print whole-genome microarrays immediately after sequence release, will facilitate transcriptional profiling in an increasing number of organisms. Transcript levels of all predicted genes can be measured simultaneously, under any given condition at several time points, to identify sets of genes whose expression levels are induced or repressed relative to a reference sample [58]. The global transcriptional profile often reflects the pathways that are directly induced or repressed in response to the primary perturbation, as well as secondary responses that might not be linked functionally to the primary perturbation.

Owing to indirect effects and genetic redundancy, the mutation of genes that are induced under a particular biological condition might not yield a specific phenotype [42]. Even though gene expression might not relate directly to protein expression [59], the proteins products of genes that are coexpressed under different conditions are often functionally related and can even interact physically with one another as part of the same pathway or complex [60–62]. Various clustering algorithms have been devised to identify coexpressed genes for functional annotation [63,64]. Because of these features, gene expression profiles have been used extensively to analyze biological perturbations. For example, a compendium of microarray gene expression profiles of yeast mutants has been used to infer the pathways affected by a mutation or a drug [65]; such compendia provide a key for interpreting how small molecules interfere with specific cellular processes (Fig. 2).

The global transcriptional regulatory network is dictated by a myriad of protein–DNA interactions and chromatin modifications. The regulation of transcription factor interactions with elements in promoter DNA nominally controls the global expression profile. Computational analysis can define potential binding sites in the promoters of co-regulated genes [66] and in alignments of promoter regions from closely related species [32]. Assignment of the cognate transcription factors to such elements remains difficult, however, probably because of the combinatorial effects between transcription factors and because their interactions with chromatin generate complex regulatory elements [41]. Indeed, such elements are only poorly predictive of co-regulation because, on average, 80% of the genes that share defined elements are not co-regulated (P. Cliften and M. Johnston, pers. commun.).

Direct analysis of protein–DNA interactions on a genome-wide scale is readily accomplished by chromatin immunoprecipitation array techniques (‘ChIP-chip’), in which DNA is crosslinked to the transcription factor of interest *in vivo* and then hybridized to a microarray [20,67]. Systematic application of this method has the potential to identify complex transcriptional regulatory circuits [20,67]. This approach can be also applied to identify any other protein–DNA interaction on a genome-wide scale, including chromatin-modifying [68,69] and DNA repair [70] complexes and replication factors [71]. Given that specificity often arises from both positively and negatively acting factors, the overlay of these datasets can prove crucial in deciphering the ultimate transcriptional hierarchy of the cell.

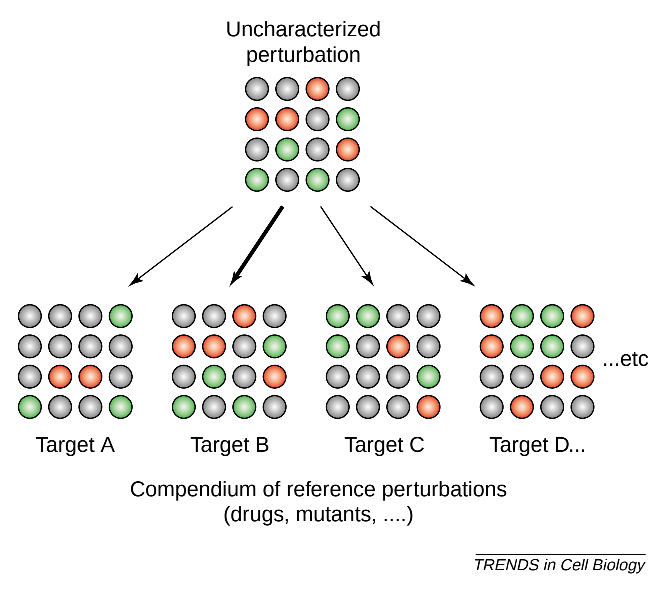

Uncharacterized perturbation
...etc
Target A
Target B
Target C
Target D...
Compendium of reference perturbations (drugs, mutants, ...)
TRENDS in Cell Biology

Fig. 2. Microarray pattern compendium. Diagram showing how a ‘compendium’ of microarray patterns, each corresponding to a different perturbation, can be used to classify an uncharacterized perturbation. Here, a hypothetical microarray with sixteen spots is shown, each measuring the expression of a single gene in a perturbed cell population relative to an unperturbed population. Black represents no change in expression, red represents induction and green represents repression. In practice, microarrays typically have thousands of spots and ratio measurements are continuous variables. Although this theoretically allows billions of different patterns, it has been estimated that in yeast there are probably several hundred discrete patterns that would result from single-gene disruptions [65].

The analysis of gene expression at various intervals after a perturbation offers the potential to computationally infer gene regulatory networks [72], their kinetics and even the protein concentration profiles of gene regulators [73]. Determining gene expression kinetics in response to numerous different perturbations can enable large-scale kinetic simulation of a gene regulation network for the cell. The integration of gene expression data with protein–protein and protein–DNA interaction networks [41,74] provides one of the first examples in which multiple data sources have been combined to deduce previously uncharted areas of the cellular map.

Protein interactions

The function of a protein is defined by the other biomolecules with which it interacts and reacts. An enormous amount of protein–protein interaction information has been obtained recently for yeast and other organisms using two-hybrid [75–77], mass spectrometry [36,78], phage-display [79] and protein fragment complementation [80] assays. Large-scale datasets derived using these methods have provided a wealth of new leads in many areas of biology. A potential difficulty with large-scale protein interaction datasets is a prevalence of false positives (interactions that are seen in an experiment but never occur in the cell or are not physiologically relevant) and false negatives (interactions that are not detected but do occur in the cell) [18,19,81,82].

Although high-quality datasets are obviously ideal, there is currently a quality/coverage tradeoff related to the speed of data acquisition. On the one hand, high-quality data are time consuming and costly to complete, leading to a low sampling of potential interactions that is biased towards known proteins. Large-scale studies, on the other hand, have a high sampling rate but can produce lower quality data. The quality of existing datasets with respect to false-positive and false-negative interactions is a complicated issue, which we discuss below. Despite these potential problems, however, protein–protein interaction networks derived from large-scale studies have proved extremely useful for defining protein function [83], examining general properties of different protein functional classes, and analyzing the topology of protein interaction networks [84].

Informatics methods can be applied to reduce the number of false positives in a dataset. By comparing datasets to benchmarks such as well-known interactions, the proportion of false positives can be estimated for a given dataset. Filtering criteria can be devised using these results combined with control data from the original experiment [36,78]. Moreover, large-scale datasets can be combined such that the overlapping set of interactions is of much higher quality than the input datasets. This has been successfully done using a simple overlap scheme [79]. This approach can be problematic if a less-sensitive dataset limits the contribution of other datasets by strict intersection (data must be in all sets).

Advanced statistical methods to combine confidence-weighted datasets should prove even more powerful [85]. If multiple datasets have low coverage and high accuracy, then a union of the sets creates a more complete dataset than an intersection. Because false positives can be reduced by dataset overlap, their occurrence is not a big problem. Instead, reducing false-negative interactions becomes a major challenge because it is extremely difficult to increase sensitivity to capture all true-positive interactions. Even for yeast, published large-scale interaction studies are far from comprehensive [18,19].

When assessing dataset quality, the definition of false positives, which can differ depending on the circumstances, can have a large effect. For instance, proteome-scale protein interaction data can be compared with the interactions derived from the crystal structures of complexes, which have arguably the highest quality of any molecular interaction data [81]. Only a small percentage of the published interaction data for yeast proteins occurring in complexes with known structures has been found to overlap with the atomic level contacts in X-ray crystal structures. But this analysis sets a very high threshold for protein interaction data because it considers interactions that are not physically direct as false positives.

When defining the function of an unknown protein that has been shown to interact with proteins of known function, an indirect interaction can be effectively used to assign functional annotation terms to the unknown molecule. Statistical methods of dataset integration to reduce false positives can be also used with information other than protein–protein interactions, such as genetic interactions, protein localizations and gene expression datasets. For instance, as mentioned above, it is known that proteins whose genes are coexpressed are more likely to interact or be part of the same complex or pathway than those whose genes are not coexpressed [60–62]. All of these data could be used together to define the reliability of specific datasets [86].

Examining patterns in network topology can prove useful for reliability assessment. Densely connected regions of a protein interaction network, which can be found computationally [19,87], often correspond to complexes that are likely to be real; for example, a six-core (a sub-network in which proteins are connected to at least six other proteins within the sub-network) from a network was predicted from phage-display-derived protein interaction motifs for Src homology domain 3 (SH3) domains in yeast and probably corresponds to an actin assembly regulatory complex [79], and a large nine-core was detected in a very large yeast network representing many interconnected complexes in the nucleolus [19].

The challenge of increasing sensitivity must be resolved through the development of wet laboratory technology. Two large-scale projects have used mass spectrometry to map protein complexes and have proved more sensitive than previous comprehensive yeast two-hybrid studies, at least as defined by a literature benchmark [36]. However, the combined results of mass spectrometry analysis still failed to recover two-thirds of the known protein associations used in a large literature-derived benchmark [19]. Interestingly, the mass spectrometry projects showed a high variability both internally and in comparison, which in part is probably due to human error and could be improved by automation and repetition. In addition, the projects used different baits for complex purification and used overexpressed versus endogenous proteins, which can have profound effects on the recovery of different protein complexes. Many different experimental methods, each with their own advantages in sampling interaction space, should be used to uncover the complete cellular interaction map.

True-negative and false-positive information from a comprehensive protein interaction screen can be useful and thus should be tracked. For example, the set of all false-positive hits derived from yeast two-hybrid screens using an SH3 domain bait might contain a subset of hits that represent direct physical interactions but might not be physiologically relevant simply because the binding partners never co-occur in the cell. Enough information can be present in this subset to derive a binding motif for the SH3 domain, similar to what can be found using phage display to screen a library of random peptides. Because this physiologically irrelevant information can have important physical meaning, it should be stored in databases along with the true-positive information such that it can be queried in the future in unforeseen ways. Tools designed to decipher ligands from interactions in this way in a fast and automated fashion must be developed in parallel with protein interaction databases. Machine-learning classification algorithms, such as the Support Vector Machine (SVM) [88], use true-positive and true-negative information to learn a decision boundary, which can be then used to classify new data. SVMs can be applied to predict protein–protein interactions but require information about proteins that are known not to interact [89].

Genetic interactions

Genetic screens for suppressors or enhancers of mutant phenotypes have been remarkably useful for identifying genes in a common pathway or process [90–92]. For example, when the phenotype of an original mutation is exacerbated by mutation of a second gene, a synthetic enhancement or, if death results, a synthetic lethal situation is scored. Tong *et al.* [23] have established a system in which a mutation in a specific query gene can be crossed to a set of 5,000 viable deletion mutants to map synthetic genetic interactions systematically. This methodology is referred to as synthetic genetic array (SGA) analysis. If the activity of a nonessential pathway is required for cellular fitness when a particular query gene is compromised functionally, then all of the components of the pathway should be identified in a comprehensive synthetic lethal screen. Thus, application of the SGA system identifies a set of synthetic genetic interactions that are enriched for the components of pathways and complexes. For example, *BIM1* encodes a protein that associates with the plus end of microtubules and participates in nuclear positioning and spindle orientation. An SGA screen with a query mutation identifies genetic interactions with kinetochore components, spindle check point proteins and the dynein–dynactin spindle orientation pathway (Fig. 3).

As the genetic network expands, complexes and pathways are expected to show a unique pattern of genetic interactions. The molecular function of previously uncharacterized genes can be thus inferred from the connectivity and the position within the network. In fact, these predictions can be precise enough to infer protein–protein interactions directly from genetic interaction data. An initial set of SGA screens suggests that many of the genes implicated in the fundamental processes required for cell division and growth show 30–50 synthetic genetic interactions, indicating that the genetic interaction map of yeast could contain over 100 000 interactions. This unexpected density of interactions indicates functional redundancy and pathway cross-talk in yeast.

As the SGA system maps interactions for deletion mutations constructed in an inbred laboratory yeast strain, perhaps the topology of the interaction network uncovered by this system can be extrapolated to more phenotypically variable outbred populations in which genetic interactions among alleles of genes presumably underlie the increased variability. Thus, large-scale genetic interaction maps created with inbred experimental systems might provide a key for deciphering the combinations of alleles underlying polygenic traits, such as human diseases, in natural populations [93]. Because gene functions are often highly conserved, a comprehensive genetic interaction map for yeast will provide a template to understand the interactions between analogous pathways in metazoans. Given the advent of RNAi technology and microarray-based transfection methodology, the SGA approach is applicable to more complex eukaryotic cells and to genetically tractable metazoan systems [49].

Protein profiling

Localization

Understanding the spatial and temporal distribution of proteins will help to define certain cell map constraints, because two proteins that interact *in vivo* must do so in the same space at the same time. Large-scale protein localization studies have been carried out in yeast by visualizing proteins either by immunofluorescence or by expressing the protein tagged to green fluorescent protein (GFP) [94]. Currently, about 54% of yeast proteins have been localized according to the Gene Ontology annotation [95] from SGD [17]. So far, genome-wide protein localization studies have not taken into account the temporal aspect of protein localization, such as the dynamic movement of proteins in and out of the nucleus [96]; however, comprehensive collections of GFP-tagged proteins should facilitate this type of analysis.

Recent advances in cryoelectron tomography that allow three-dimensional (3D) visualization of the actin cytoskeleton and the 26S proteasome in *Dictyostelium* cells forecast the ability to take a 3D snapshot of the structure of a cellular proteome at a resolution of less than 2 nm [97]. The dynamic analysis of protein localization will obviously become more complex as large-scale studies move from yeast to multicellular organisms, which depend on the regulation of protein localization for cellular differentiation during development.

Identification

Advances in mass spectrometry have led to fast and accurate protein identification, as long as the protein

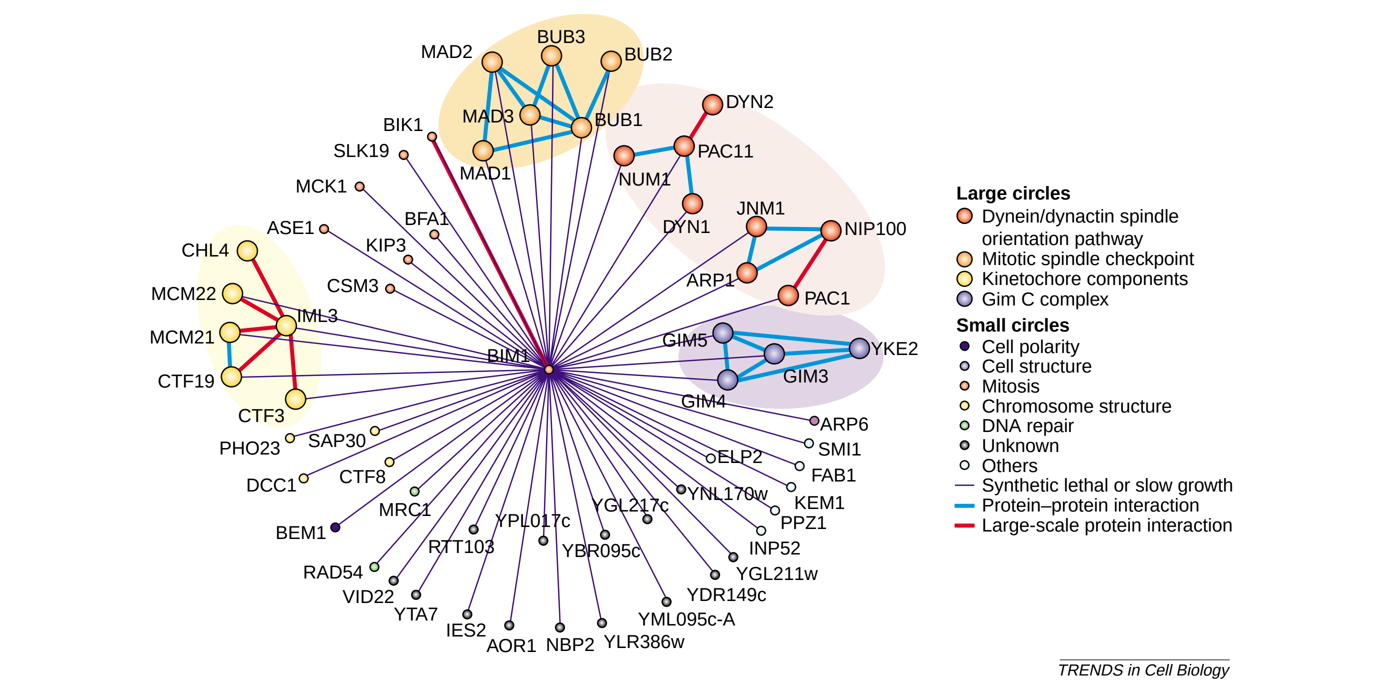


Fig. 3. Integration of genetic and protein interactions. Shown is a set of synthetic lethal and slow growth interactions derived from an SGA screen with a *BIM1* query originally from the SGA study of Tong *et al.* [23]. It is clear that genetic interaction data, specifically synthetic lethal and slow growth interactions, are enriched for proteins that physically interact with each other or are in the same complex or pathway. All genes on this map are non-essential genes, as is normally the case with the SGA technique. Gene annotation is based on the Gene Ontology terms in the SGD database. Annotation of the interactions is based on those in the BIND database [12].

already exists uniquely in a sequence database [98]. Whether a protein is present or not in a sample can be used to map signaling pathways, complexes [36,78] and even all of the proteins in an organelle [99]. One of the next frontiers in this field is the ability to measure quantities of proteins in the cell. Genome-scale protein quantification is not yet feasible, but methods for determining relative levels of protein between samples have been developed [100]. Alternatively, arrays of cell colonies, each expressing a different fluorescent tagged protein, should enable quantification of protein expression in response to specific genetic and environmental perturbations [101]. A measurement of the levels of all proteins in a cell over time will provide insight into the molecular basis of different cellular states – a prerequisite for their modeling.

Post-translational modification mapping

Protein regulation by means of post-translational modifications (PTMs) can determine when and where a protein is active in the cell, and mass spectrometry and protein chips are being applied to systematically identify PTMs in a proteome. Mass spectrometry holds great promise for proteome-wide PTM mapping: the large-scale mapping of phosphorylation sites has been performed for yeast [102,103], and a technique based on mass spectrometry for mapping *O*-linked *N*-acetylglucosamine PTMs has been developed recently [104]. But the wide range of protein modifications from acetylation to lipid modification will be problematic to overcome [103,105].

Biochemical approaches to PTM discovery also exist. For example, protein chips that display a whole proteome on a relatively small surface for functional testing in different assays [106] offer the potential to identify all possible targets for a particular kinase and, therefore, to identify a global phosphoprotein regulatory map including all kinases and their substrates.

Structure

Structural genomics projects ([107,108]; and see PSB Structural Genomics: http://www.rcsb.org/pdb/strucgen.html) have the potential to define the 3D structure of all proteins, generally by X-ray crystallography, but whether this goal can be achieved in a high-throughput manner is still controversial [109]. Almost the whole crystallography process can be automated from protein expression, to crystallization trials, to positioning the sample in a synchotron X-ray beam line. If the crystal structure is good enough, even the final structural modeling step can be done computationally. But bottlenecks still arise in this approach during the protein expression and crystallization step, especially for eukaryotic proteins that are difficult to express. Crystallizing membrane proteins and proteins that are structured only when part of a physiological complex [110] still represent tough challenges.

Homology modeling techniques can generally compute the structure of a protein if the structure of another protein with greater than 30% sequence identity is known [111]. Thus, if one protein species cannot be crystallized easily, another with a similar sequence can be attempted. It has been suggested that roughly 16 000 carefully chosen protein structure targets could cover the structural diversity of most known proteins [112]. Targets from among this reasonably sized set could be chosen in an order that provides maximum information rapidly. For example, proteins that are involved in a cellular module of interest, such as a complex or an organelle, could be chosen first and the module investigated before completion of the whole structural genomics project.

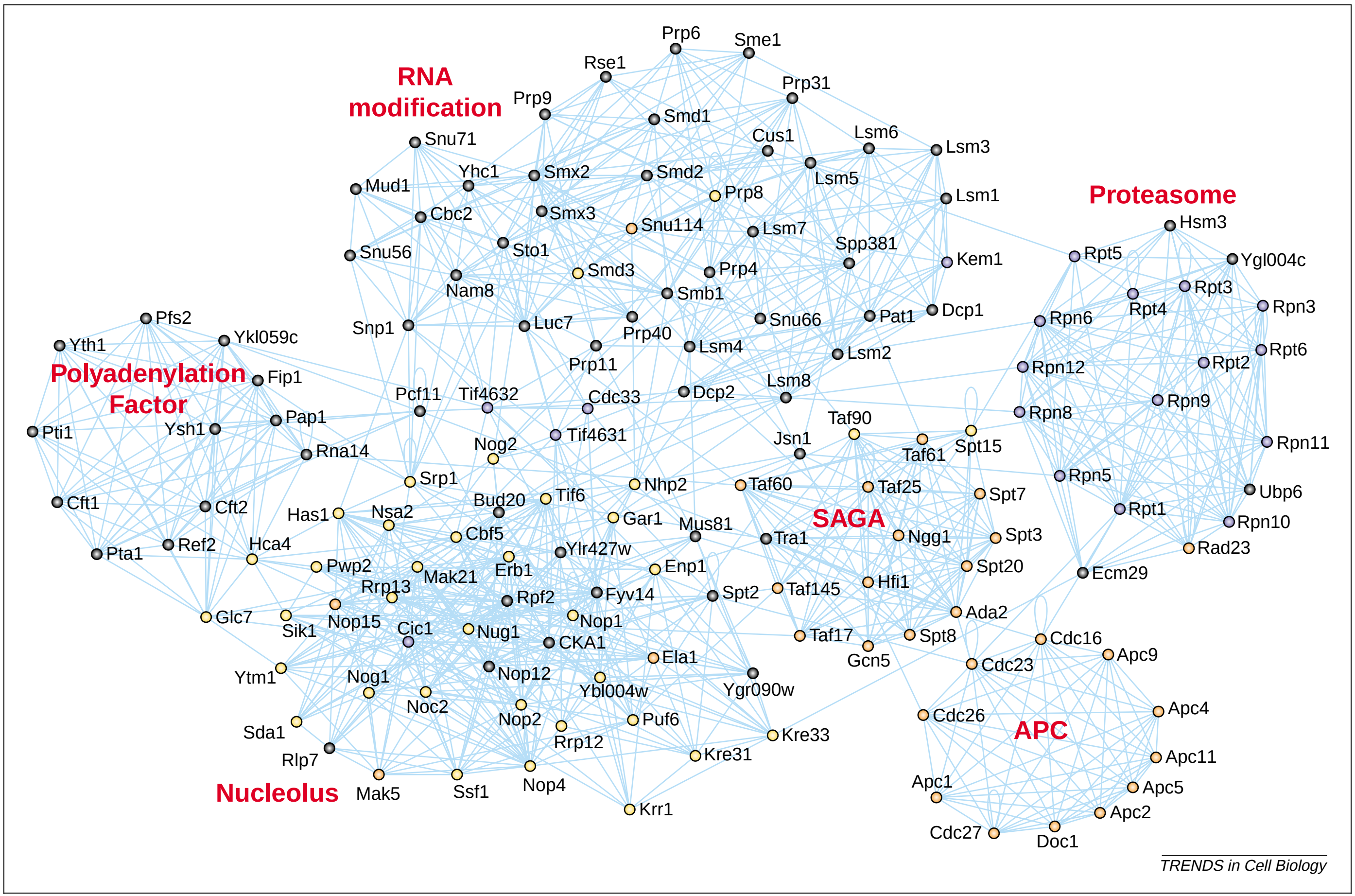
Enzymatic function

On a molecular level, proteins have many different enzymatic and ligand-specific binding functions, each with their own kinetic and thermodynamic properties. Protein functional assays have been developed to study these protein properties on a large scale. For example, the complete set of yeast genes has been expressed as proteins tagged to glutathione *S*-transferase (GST) and affinity-purified to assay for enzymatic functions that are known to occur but remain to be linked to a catalytic protein or complex [113]. Specific protein functions, such as protein kinase activity [114], have been assayed in nanoliter-sized wells on a large scale. Kinetic rate constants for protein-catalyzed chemical reactions also must be measured on a large scale, and this is planned at least for *Escherichia coli* (Project CyberCell: http://www.projectcybercell.com/). The results of such studies will provide detailed information that will eventually allow kinetic simulations, or models, of biological systems [115].

Discovery by mining functional genomics databases

The collection of large-scale functional genomics data in yeast has led already to some fundamental insights about biological networks and gene function. In a first example, a combination of genome-wide transcriptional profiles, large-scale protein–protein interaction mapping and phenomic analysis has identified a large group of co-regulated genes, called the 'RiBi regulon', that participates in ribosome biogenesis [116]. This co-regulated set contains more than 200 uncharacterized genes, nearly half of which are essential for viability [64]. Two strongly predicted potential binding sites, termed PAC and RRPE, lie upstream of most of these genes [66]. Sfp1 emerged unexpectedly as a candidate transcription factor for the RiBi regulon from a systematic screen for yeast mutants that prematurely commit to cell division and display a small cell size [41]. Unbiased computational clustering of all known protein interactions identified a large previously unknown set of related complexes composed of many of the same nucleolar proteins [19] (Fig. 4), many of which have been since assigned to discrete steps in either 40S or 60S ribosomal particle biogenesis [117,118]. These data suggest that 30% or more of all essential yeast genes might be dedicated to the processing of noncoding RNA.

In a second example of functional genomic insight, it has been shown that the connectivity distribution of protein interaction networks follows a power law [119,120]; that is, a few proteins called 'hubs' are involved in many interactions, whereas many proteins are involved in a few interactions. Evolution can generate such highly connected hubs by building successive layers of regulatory factors onto essential cellular processes. Importantly,



TRENDS in Cell Biology

Fig. 4. Clusters of highly connected nuclear protein complexes. The central densest region of a large interaction network containing over 15 000 protein interactions from yeast is shown. The interactions were collected from all large-scale studies done to date, as well as the MIPS [16] and BIND [12] databases. Known molecular complexes can be seen clearly, as well as a large, previously unsuspected nucleolar complex. All of the proteins in this network are connected to all other proteins in the network by at least nine interactions. Proteins are colored by cellular localization, as defined in the Gene Ontology terms in the SGD database. In 1000 randomly permuted networks, the mean highest *k*-core (see text) was 7 (s.d. = 0), indicating that a nine-core is highly unlikely to occur by chance. This analysis was done in Ref. [19]. The 19S regulatory subunit of the proteasome, which is involved in targeted protein degradation, is labeled 'proteasome'. APC, anaphase-promoting complex; SAGA, Spt-Ada-Gcn5-acetyltransferase (SAGA) transcriptional activator/histone acetyltransferase complex.

power law networks are robust against random attack (protein deletion). In simulated attacks, such networks stay statistically coherent until most of the protein nodes are eliminated. This property is biologically attractive because it can help to explain how evolution can create systems that are buffered from the wide-ranging effects of random mutations. If the highly connected hubs are removed first, however, the network quickly disintegrates into disconnected components. This fact is also biologically relevant because it has been shown that hubs in a power law network are more likely to be essential genes [119]. Consideration of statistical network properties has many practical ramifications for rational drug design and our understanding of evolved disease states, such as cancer.

As a final example of the discovery value in large-scale datasets, we have explored the global connectivity of a protein functional class. Of an integrated network of more than 15 000 yeast protein–protein interactions, we extracted those involving only kinases and phosphatases (~170 proteins). Interestingly, these signaling molecules are assembled into a highly connected network (Fig. 5), an observation originally noted by Ho *et al.* [36]. This finding reflects an unusual property because proteins in other functional classes, such as the set of about 180 transcription factors in the MIPS database (Fig. 5) and similar-sized sets of random proteins, do not form highly connected networks. Thus, protein–protein interaction studies focused specifically on the kinases and phosphatases should efficiently chart the basic signaling circuitry of an organism and provide a scaffold for linking together all cellular processes regulated by protein phosphorylation.

Databases and visualization

Building an accurate and complete cellular map, tantamount to a dynamic high dimensional information matrix, will require the integration of many layers of systematic cell and molecular biology and many direct lines of research. To this end, many approaches are possible. Some groups, such as the Alliance for Cell Signaling [121], have undertaken to map pathways in specific cells (initially lymphocytes and cardiac myocytes) by vertically integrating systematically derived data from many member laboratories. Smaller groups are attacking a single model organism, using either a single specific technique such as RNAi [49] or multiple orthogonal techniques such as protein interaction mapping and expression data [122].

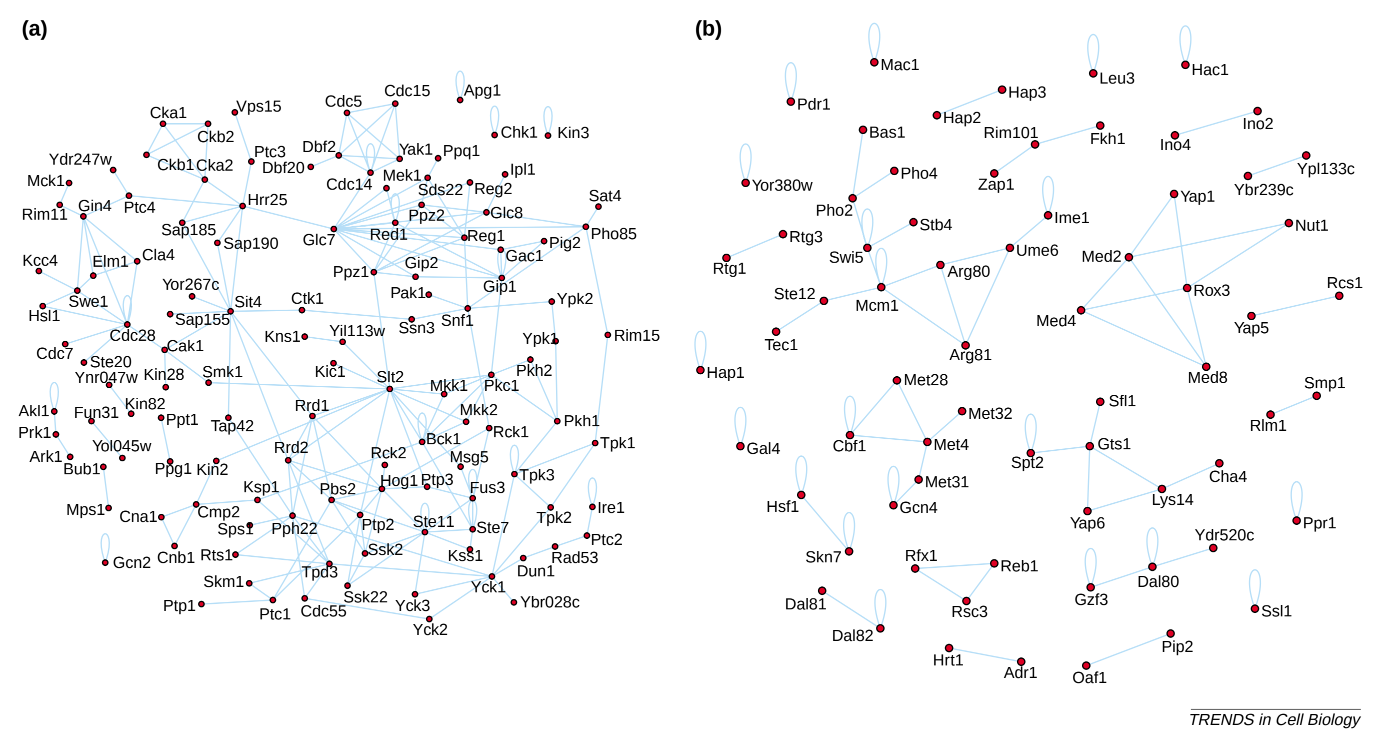


Fig. 5. A large network of protein–protein interactions among kinases and phosphatases in yeast. (a) Kinases and phosphatases are very well connected in a large protein–protein interaction network. (b) Transcription factors, a functional class similar in size to the kinase/phosphatase class, are not. This is an example of an unanticipated result that is completely unobtainable without genome-wide studies. Loops indicate self-interactions.

A comprehensive multidimensional cell map would require, in principle, full dynamic knowledge of all parts of the cell in time and space [123], including direct physical interactions, precisely delineated binding sites, kinetics and reaction rates as well as biomolecular concentrations (protein, RNA and small molecule) at all stages of the cell cycle and in all differentiated states with all genetic interactions, and so on. Whether useful information of such complexity can be even acquired remains to be seen. Even a limited subset of these data will require powerful information storage, query and analysis engines to handle data manipulation computationally. Current representational models of pathways and cell simulation will need to evolve substantially to manage these data meaningfully.

Databases such as BIND [12], DIP [13], MINT [14], GRID [15], SGD [17] and MIPS [16] are intended to serve as a repository for protein and genetic interactions and associated regulatory events, as occur in cell signaling. Gene expression databases already store huge amounts of DNA microarray information from many organisms [124,125], and yet other databases can store transcription factor [126], metabolic pathway [29,127,128] and gene regulatory network [129] data. Building and maintaining a high-quality database requires a substantial amount of effort. Thus, creating a database large enough to capture cell map information will require massive community investment and commitment, ranging from the individual researcher to the funding agency and journals, as well as innovation from database developers. Pathway simulation engines [4,115] are available to examine quantitatively mathematical models of these data.

All of this must be tied together using data standards (see BioPax: Biological Pathways Exchange: http://www.biopax.org; Proteomics Standards Initiative: http://psidev.sourceforge.net/; Systems Biology Markup Language: http://sbw-sbml.org/) and Web services that can be easily queried for information [130,131]. Machine learning tools such as SVMs [132], Bayesian nets [133] and decision trees [134] will be required to integrate, to filter and to recognize patterns automatically in this enormous multidimensional dataset, and the use of network visualization and modeling tools such as Cytoscape (see http://www.cytoscape.org), Osprey and BioLayout [15,74,135,136] will be necessary to understand data relationships quickly and to make biologically relevant predictions. Indeed, these visualization tools must be developed as the interactive entry point to the integrated cell map, where a gene of interest connects directly to the latest information about that gene and its relationships.

As an example, the initial version of Cytoscape can represent several concurrent aspects of the multidimensional cell map. Figure 6 shows two versions of the protein kinase C (PKC) pathway from yeast. A manually constructed version represents a limited connection map of PKC pathway proteins [137], whereas a version automatically constructed by Cytoscape is based on a data file containing a large set of interactions between proteins, genes and transcription factors, combined with original microarray gene expression data from Roberts *et al.* [137]. In addition to representing data associated with the PKC pathway more fully, the Cytoscape network can be queried interactively to reveal several layers of information, which can be crucial for hypothesis generation. Discoveries

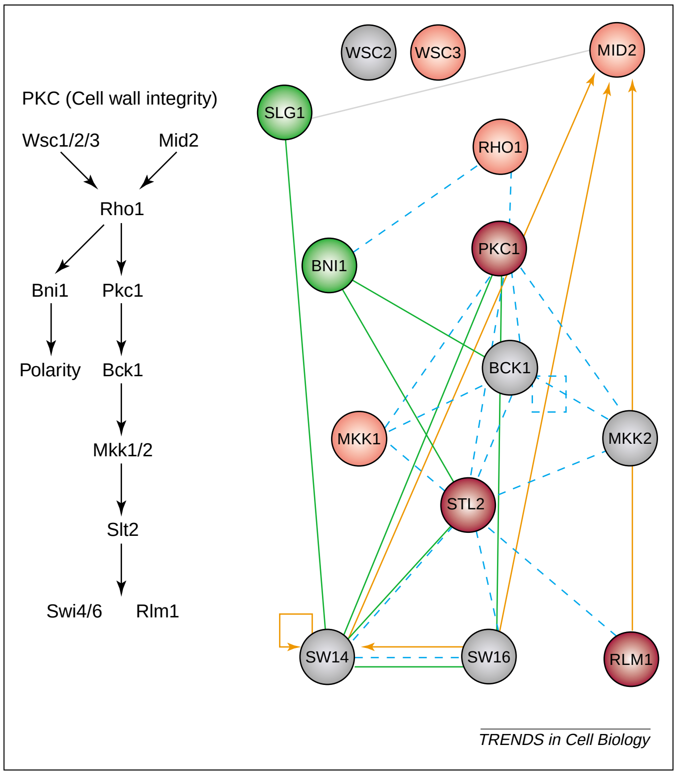


Fig. 6. Visualizing a part of the dynamic multidimensional cell map. Left, a reproduction of the PKC pathway in yeast manually drawn in *c.*2000 [137]. Right, the same proteins laid out automatically using protein–protein, protein–DNA and genetic interactions overlaid with gene expression data from [137] using the Cytoscape tool. The input to Cytoscape was a data file containing over 15 000 protein–protein interactions, over 300 genetic interactions and more than 5,600 protein–DNA (transcription factor) interactions. Over 5500 proteins are represented. Thus, the PKC pathway is just one of many that can be automatically and interactively visualized. Interestingly, the *WSC2* and *WSC3* genes that are connected to the pathway in the manual drawing are not connected in the Cytoscape drawing, indicating that databases are still missing important interactions. Green lines indicate synthetic sick or synthetic lethal interactions from the SGA screen of Tong *et al.* [23], yellow arrows point from a transcription factor to a regulated gene, and blue broken arrows represent protein–protein interactions. Circles indicate genes or their cognate proteins and are colored red or green according to whether their expression was upregulated or downregulated, respectively, as compared with wild-type S288c yeast cells, when PKC was overexpressed as a dominant-activated protein (R398A mutation).

prompted by large-scale datasets generated across all manner of model systems will depend on data assembly tools such as Cytoscape [15,74,135].

Perhaps one of the most powerful aspects of large-scale studies is the potential for comprehensive analysis. Completeness of functional knowledge of the cell is an ultimate goal but will obviously be difficult to achieve. Classical research clusters within certain fields, and thus only expands knowledge at the field periphery. Also, classical research tends to focus only on fashionable fields, leaving older or less trendy fields without much innovation. As can be seen in Fig. 1, for instance, much of yeast research has focused on a subset of genes that the community finds particularly interesting, such as those involved in cell cycle regulation or chromosome dynamics. In fact, only a few papers are being currently published on metabolism (Fig. 1), and yet its integration with cell regulation pathways is vital for a complete cell map. For all organisms there are examples of genes about which no information is known from any method. This class of uncharacterized genes has been called the 'Unknome' [138]. Obviously, achieving a more even distribution of functional categories in large-scale studies would be the first approach to reducing the size of the Unknome.

Concluding remarks

As high-throughput functional genomics and proteomics technology and bioinformatics develop concurrently, they will become more accessible to the individual laboratory, which will be thus empowered to ask increasingly more interesting biological questions. For example, many proteins are highly conserved across evolution, and it will be interesting to determine the extent to which the cell map is conserved. All aspects of evolution that have been studied at the sequence level can be also studied at the cell map level, but this requires data across species. This should enable us to understand the evolution of complex features in humans by network differentiation and evolution from simpler systems. Furthermore, the cell map will facilitate large-scale modeling of the cell, although building computational systems that have enough highly detailed information and computer processing resources for a complete cell model will probably take many decades. Only the tight integration of wet-laboratory biology and bioinformatics will enable us to overcome these challenges.

Acknowledgements

We thank: Yuen Ho, Chris Hogue, Debbie Marks, Tony Pawson, Chris Sander and other colleagues for helpful discussion; Mark Robinson for performing the analysis shown in Fig. 1; Trey Ideker for assistance with Fig. 6; Ronald Jansen for critically reading the manuscript; and Mike Moran and Daniel Figeys from MDSP for help with the kinase–phosphatase map.

References

1 Fields, S. (2001) Proteomics. Proteomics in genomeland. *Science* 291, 1221–1224

2 Fields, S. and Johnston, M. (2002) Genomics. A crisis in postgenomic nomenclature. *Science* 296, 671–672

3 Steven, W.H. *et al.* (2003) Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol.* 13, 43–50

4 Covert, M.W. *et al.* (2001) Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.* 26, 179–186

5 Cross, F.R. *et al.* (2002) Testing a mathematical model of the yeast cell cycle. *Mol. Biol. Cell* 13, 52–70

6 Hartwell, L.H. *et al.* (1999) From molecular to modular cell biology. *Nature* 402, C47–C52

7 Ideker, T. *et al.* (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2, 343–372

8 Levsky, J.M. and Singer, R.H. (2003) Gene expression and the myth of the average cell. *Trends Cell Biol.* 13, 4–6

9 Grunenfelder, B. and Winzeler, E.A. (2002) Treasures and traps in genome-wide data sets: case examples from yeast. *Nat. Rev. Genet.* 3, 653–661

10 Oliver, S.G. *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46

11 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 546, 563–567

12 Bader, G.D. *et al.* (2001) BIND – the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 29, 242–245

13 Xenarios, I. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305

14 Zanzoni, A. *et al.* (2002) MINT: a Molecular INTeraction database. *FEBS Lett.* 513, 135–140

15 Breitkreutz, B.J. *et al.* (2002) The GRID: the general repository for interaction datasets. *Genome Biol.* 3, R0013

16 Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34

17 Dwight, S.S. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* 30, 69–72

18 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403

19 Bader, G.D. and Hogue, C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* 20, 991–997

20 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804

21 Horak, C.E. *et al.* (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* 16, 3017–3033

22 Iyer, V.R. *et al.* (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409, 533–538

23 Tong, A.H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364–2368

24 Endy, D. and Brent, R. (2001) Modelling cellular behaviour. *Nature* 409 (Suppl.), 391–395

25 Edwards, J.S. and Palsson, B.O. (2000) Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics* 1, 1

26 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

27 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

28 Benson, D.A. *et al.* (2002) GenBank. *Nucleic Acids Res.* 30, 17–20

29 Overbeek, R. *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28, 123–125

30 Paley, S.M. and Karp, P.D. (2002) Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics* 18, 715–724

31 Mathe, C. *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103–4117

32 Cliften, P.F. *et al.* (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* 11, 1175–1186

33 Kumar, A. *et al.* (2002) An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* 20, 58–63

34 Shoemaker, D.D. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature* 409, 922–927

35 Lipton, M.S. *et al.* (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11049–11054

36 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183

37 Vaglio, P. *et al.* (2003) WorfDB: the *Caenorhabditis elegans* ORFeome Database. *Nucleic Acids Res.* 31, 237–240

38 Paigen, K. and Eppig, J.T. (2000) A mouse phenome project. *Mamm. Genome* 11, 715–717

39 Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906

40 Chang, M. *et al.* (2002) A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16934–16939

41 Jorgensen, P. *et al.* (2002) Systematic identification of pathways that couple cell growth and division in yeast. *Science* 297, 395–400

42 Giaever, G. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391

43 Page, N. *et al.* (2003) A *Saccharomyces cerevisiae* genome-wide mutant screen for altered sensitivity to K1 killer toxin, *Genetics* 163, 875–894

44 Ni, L. and Snyder, M. (2001) A genomic study of the bipolar bud site selection pattern in *Saccharomyces cerevisiae* . *Mol. Biol. Cell* 12, 2147–2170

45 Bonangelino, C.J. *et al.* (2002) Genomic screen for vacuolar protein sorting genes in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* 13, 2486–2501

46 Jiang, B. *et al.* (2002) Novel strategies in antifungal lead discovery. *Curr. Opin. Microbiol.* 5, 466–471

47 Steinmetz, L.M. *et al.* (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.* 31, 400–404

48 Piano, F. *et al.* (2000) RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Curr. Biol.* 10, 1619–1622

49 Kamath, R.S. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421, 231–237

50 Hudson, D.F. *et al.* (2002) Reverse genetics of essential genes in tissue-culture cells: 'dead cells talking'. *Trends Cell Biol.* 12, 281–287

51 Nolan, P.M. *et al.* (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat. Genet.* 25, 440–443

52 Wu, R.Z. *et al.* (2002) Cell-biological applications of transfected-cell microarrays. *Trends Cell Biol.* 12, 485–488

53 Kononen, J. *et al.* (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* 4, 844–847

54 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470

55 Packer, A. (ed.) (2002) Nature Genetics Chipping Forecast II. *Nat. Genet.* 32 (Suppl.), 461–552

56 Hughes, T.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19, 342–347

57 Nuwaysir, E.F. *et al.* (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* 12, 1749–1755

58 Ashby, M. and Rine, J. (10-29-1996) Methods for drug screening. The Regents of the University of California. Oakland, CA. USA., Patent number: 5,569,588

59 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–934

60 Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 29, 3513–3519

61 Ge, H. *et al.* (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29, 482–486

62 Jansen, R. *et al.* (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.* 12, 37–46

63 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

64 Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255–265

65 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126

66 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285

67 Horak, C.E. and Snyder, M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* 350, 469–483

68 Robyr, D. *et al.* (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* 109, 437–446

69 van Leeuwen, F. *et al.* (2002) Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* 109, 745–756

70 Tanaka, T. and Nasmyth, K. (1998) Association of RPA with chromosomal replication origins requires an Mcm protein, and is regulated by Rad53, and cyclin- and Dbf4-dependent kinases. *EMBO J.* 17, 5182–5191

71 Blat, Y. and Kleckner, N. (1999) Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* 98, 249–259

72 Pe'er, D. *et al.* (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 (Suppl. 1), S215–S224

73 Ronen, M. *et al.* (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U. S. A.* 99, 10555–10560

74 Ideker, T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 (Suppl. 1), S233–S240

75 Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627

76 Uetz, P. (2002) Two-hybrid arrays. *Curr. Opin. Chem. Biol.* 6, 57–62

77 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574

78 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147

79 Tong, A.H. *et al.* (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295, 321–324

80 Remy, I. *et al.* (2002) Detection and visualization of protein interactions with protein fragment complementation assays. *Methods Mol. Biol.* 185, 447–459

81 Edwards, A. *et al.* (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* 18, 529

82 Kemmeren, P. *et al.* (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* 9, 1133–1143

83 Schwikowski, B. *et al.* (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261

84 Albert, R. *et al.* (2000) Error and attack tolerance of complex networks. *Nature* 406, 378–382

85 Gerstein, M. *et al.* (2002) Proteomics. Integrating interactomes. *Science* 295, 284–287

86 Deane, C.M. *et al.* (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* 1, 349–356

87 Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2

88 Gaasterland, T. and Bekiranov, S. (2000) Making the most of microarray data. *Nat. Genet.* 24, 204–206

89 Bock, J.R. and Gough, D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics* 17, 455–460

90 Guarente, L. (1993) Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet.* 9, 362–366

91 Novick, P. *et al.* (1989) Suppressors of yeast actin mutations. *Genetics* 121, 659–674

92 Forsburg, S.L. (2001) The art and design of genetic screens: yeast. *Nat. Rev. Genet.* 2, 659–668

93 Hartman, J.L. *et al.* (2001) Principles for the buffering of genetic variation. *Science* 291, 1001–1004

94 Kumar, A. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.* 16, 707–719

95 The Gene Ontology Consortium, (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29

96 Shimada, Y. *et al.* (2000) Nuclear sequestration of the exchange factor Cdc24 by Far1 regulates cell polarity during yeast mating. *Nat. Cell Biol.* 2, 117–124

97 Medalia, O. *et al.* (2002) Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* 298, 1209–1213

98 Mann, M. *et al.* (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473

99 Andersen, J.S. *et al.* (2002) Directed proteomic analysis of the human nucleolus. *Curr. Biol.* 12, 1–11

100 Smolka, M. *et al.* (2002) Quantitative protein profiling using two-dimensional gel electrophoresis, isotope-coded affinity tag labeling, and mass spectrometry. *Mol. Cell Proteomics.* 1, 19–29

101 Dimster-Denk, D. *et al.* (1999) Comprehensive evaluation of isoprenoid biosynthesis regulation in *Saccharomyces cerevisiae* utilizing the genome reporter matrix. *J. Lipid Res.* 40, 850–860

102 Oda, Y. *et al.* (2001) Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* 19, 379–382

103 Ficarro, S.B. *et al.* (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 20, 301–305

104 Wells, L. *et al.* (2002) Mapping sites of *O*-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Mol. Cell Proteomics* 1, 791–804

105 Mann, M. *et al.* (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol.* 20, 261–268

106 Grayhack, E.J. and Phizicky, E.M. (2001) Genomic analysis of biochemical function. *Curr. Opin. Chem. Biol.* 5, 34–39

107 Christendat, D. *et al.* (2000) Structural proteomics of an archaeon. *Nat. Struct. Biol.* 7, 903–909

108 Stevens, R.C. *et al.* (2001) Global efforts in structural genomics. *Science* 294, 89–92

109 Montelione, G.T. (2001) Structural genomics: an approach to the protein folding problem. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13488–13489

110 Dyson, H.J. and Wright, P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12, 54–60

111 Pieper, U. *et al.* (2002) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* 30, 255–259

112 Vitkup, D. *et al.* (2001) Completeness in structural genomics. *Nat. Struct. Biol.* 8, 559–566

113 Martzen, M.R. *et al.* (1999) A biochemical genomics approach for identifying genes by the activity of their products. *Science* 286, 1153–1155

114 Zhu, H. *et al.* (2000) Analysis of yeast protein kinases using protein chips. *Nat. Genet.* 26, 283–289

115 Slepchenko, B.M. *et al.* (2002) Computational cell biology: spatiotemporal simulation of cellular events. *Annu. Rev. Biophys. Biomol. Struct.* 31, 423–441

116 Gasch, A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257

117 Dragon, F. *et al.* (2002) A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* 417, 967–970

118 Fatica, A. and Tollervey, D. (2002) Making ribosomes. *Curr. Opin. Cell Biol.* 14, 313–318

119 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

120 Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18, 1283–1292

121 Gilman, A.G. *et al.* (2002) Overview of the alliance for cellular signaling. *Nature* 420, 703–706

122 Walhout, A.J. *et al.* (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* 12, 1952–1958

123 Vidal, M. (2001) A biological atlas of functional maps. *Cell* 104, 333–339

124 Edgar, R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210

125 Gollub, J. *et al.* (2003) The Stanford microarray database: data access and quality assessment tools. *Nucleic Acids Res.* 31, 94–96

126 Matys, V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 374–378

127 Karp, P.D. *et al.* (2002) The EcoCyc database. *Nucleic Acids Res.* 30, 56–58

128 van Helden, J. *et al.* (2001) From molecular activities and processes to biological function. *Brief. Bioinform.* 2, 81–93

129 Ananko, E.A. *et al.* (2002) GeneNet: a database on structure and functional organisation of gene networks. *Nucleic Acids Res.* 30, 398–401

130 Michalickova, K. *et al.* (2002) SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics* 3, 32

131 Stein, L. (2002) Creating a bioinformatics nation. *Nature* 417, 119–120

132 Furey, T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914

133 Drawid, A. and Gerstein, M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* 301, 1059–1075

134 Bertone, P. *et al.* (2001) SPINE: an integrated tracking database and

data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* 29, 2884–2898

135 Breitkreutz, B.J. *et al.* (2002) Osprey: a network visualization system. *Genome Biol.* 3, R0012

136 Enright, A.J. and Ouzounis, C.A. (2001) BioLayout – an automatic graph layout algorithm for similarity visualization. *Bioinformatics* 17, 853–854

137 Roberts, C.J. *et al.* (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880

138 Greenbaum, D. *et al.* (2001) Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.* 11, 1463–1468

Have you seen our 'Tube Morphogenesis' series, which began in the August 2002 issue?

Articles published to date:

Tube Morphogenesis (Editorial)
Mark A. Krasnow and W. James Nelson (2002)
Trends Cell Biol. 12, 251

Tubulogenesis in the developing mammalian kidney
Gregory R. Dressler (2002)
Trends Cell Biol. 12, 390–395

Vascular cell biology in vivo: a new piscine paradigm?
Brant M. Weinstein (2002)
Trends Cell Biol. 12, 439–445

Tubes and the single *C. elegans* excretory cell
Matthew Buechner (2002)
Trends Cell Biol. 12, 479–484

Extracellular matrix in vascular morphogenesis and disease: structure versus signal
Benjamin S. Brooke, Satyajit K. Karnik and Dean Y. Li (2003)
Trends Cell Biol. 13, 51–56

Branching morphogenesis of the lung: new molecular insights into an old problem
Andrew P. McMahon and Pao-Tien Chuang (2003)
Trends Cell Biol. 13, 86–91

Making vascular networks in the adult: branching morphogenesis without a roadmap
Yuval Dor, Valentin Djonov and Eli Keshet (2003)
Trends Cell Biol. 13, 131–136

Epithelial polarity and tubulogenesis *in vitro*
Mirjam M.P. Zegers *et al.(2003)*
Trends Cell Biol. 13, 169–177

Constructing an organ: the *Drosophila* salivary gland as a basic model for tube formation
Elliott Abrams, Melissa Vining and Deborah Andrews
Trends Cell Biol. 13, 247–254

Drosophila tracheal morphogenesis: intricate cellular solutions to basic plumbing problems
Christos Samakovlis *et al.* (June 2003, in press)

How to make tubes: signaling by the c-Met receptor tyrosine kinase
Walter Birchmeier and Marta Rosario (June 2003, in press)

Making a zebrafish kidney: a tale of two tubes
Iain Drummond (July 2003, this issue)

Other reviews planned for the series:

Tubulogenesis in *Drosophila* and mammalian kidney development
Helen Skaer

Role of polycystic kidney disease protein in establishing and maintaining tubular structure
Alessandra Boletta and Greg Germino

Epimorphin and mammary gland tubulogenesis
Derek Radisky and Mina Bissell