Computational Prediction of Protein-Protein Interactions

Anton J. Enright[1], Lucy Skrabanek[2] and Gary D. Bader[1]

[1] Computational Biology Center, Memorial Sloan-Kettering Cancer Center,
307 East 63rd Street, New York, NY 10021, USA.
Email: **enright@cbio.mskcc.org**, **bader@cbio.mskcc.org** Ph: (646) 735 8079, Fax: (646) 735 0021

[2] Department of Physiology and Biophysics and Institute for Computational Biomedicine,
Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021, USA.
Email: **las2017@med.cornell.edu** Ph: (212) 327 7361, Fax: (212) 860 3369

Abstract

Recently a number of computational approaches have been developed for the prediction of protein-protein interactions. Complete genome sequencing projects have provided the vast amount of information needed for these analyses. These methods utilize the structural, genomic and biological context of proteins and genes in complete genomes to predict protein interaction networks and functional linkages between proteins. Given that experimental techniques remain expensive, time-consuming and labor-intensive, these methods represent an important advance in proteomics. Some of these approaches utilize sequence data alone to predict interactions, while others combine multiple computational and experimental datasets to accurately build protein interaction maps for complete genomes. These methods represent a complementary approach to current high-throughput projects whose aim is to delineate protein interaction maps in complete genomes. In this chapter, we will describe a number of computational protocols for protein interaction prediction based on the structural, genomic and biological context of proteins in complete genomes, and detail methods for protein interaction network visualization and analysis.

Keywords: Genome context, gene fusion, phylogenetic profiles, gene neighborhood, protein interaction networks, visualization.

1. Introduction

One of the current goals of proteomics is to map the protein interaction networks of a large number of model organisms ***(1)***. Protein-protein interaction information allows the function of a protein to be defined by its position in a complex web of interacting proteins. Access to such information will greatly aid biological research and potentially make the discovery of novel drug targets much easier. Previously the detection of protein-protein interactions was limited to labor-intensive experimental techniques such as co-immunoprecipitation or affinity chromatography. High-throughput experimental techniques such as yeast two-hybrid and mass spectrometry have now also become available for large-scale detection of protein interactions. These methods however, may not be generally applicable to all proteins in all organisms, and may also be prone to systematic error. Recently, a number of complementary computational approaches have been developed for the large-scale prediction of protein-protein interactions based on protein sequence, structure and evolutionary relationships in complete genomes.

Initially computational prediction of protein-protein interactions was strictly limited to proteins whose three-dimensional structures had been determined. These methods predicted protein-protein interaction based on the *structural context* of proteins. Recent advances in complete genome sequencing have however provided a wealth of genomic information. It is now possible to establish the *genomic context* of a given gene in a complete genome ***(2,3)***. A gene is no longer thought of as a single protein-coding entity but as part of a coordinated network of interacting proteins. The potential for two proteins to interact is not only specified by the physical and structural properties of their structures, but is also encoded at a genomic level. For example, interacting genes are generally co-expressed ***(4-6)*** (both temporally and spatially). In other words, the fact that two proteins have the physical potential to interact is meaningless unless they are present in the same part of the cell at the same time. Other examples of genomic context include the co-localization of genes on chromosomes, the complete fusion of pairs of genes, correlated mutations between interacting protein families, and phylogenetic gene profiles. Even in the absence of structural or sequence information, one can detect the evolutionary fingerprints of pairs of interacting proteins from their genomic context. A number of these computational approaches also take advantage of high-throughput experimental information such as gene-expression data, cellular locality and molecular complex information ***(7,8)***. These hybrid computational approaches exploit both the genomic and *biological context* of genes and proteins in complete genomes in order to predict interactions.

In this chapter, we will describe computational methods and resources available for protein-protein interaction prediction that exploit the structural, genomic and biological contexts of proteins in complete genomes. In addition to algorithms and methods for interaction prediction, a number of useful databases pertaining to protein-protein interaction will be described. These databases combine a large amount of data from both computational and experimental techniques. Finally, a number of tools for protein interaction network visualization and analysis will be described. Methods are presented in historical order together with online access information. Where available, detailed computational protocols will also be provided for each method.

1.1 Structural Context Approaches

Computational prediction of protein-protein interactions consists of two main areas (i) the mapping of protein-protein interactions i.e., determining whether two proteins are likely to interact, and (ii) the understanding of the mechanism of protein-protein interactions and the identification of residues in proteins which are involved in those interactions. The first successful computational analyses of protein-protein interactions, used the structural context of proteins in order analyze known protein interaction interfaces in order to determine physical rules determining protein-protein interaction specificity. Unlike other computational methods that use an evolutionary or genomic context to predict interaction, structural approaches tend to be more limited in terms of scale, as only a small proportion of protein sequences have accurate three-dimensional structures deposited in the Protein Databank (PDB) ***(9)***. However, structural approaches allow for a much more detailed analysis of protein interactions than the genome-context based approaches. Structural approaches can determine, not only whether two proteins interact, but also the physical characteristics of the interaction, and residues (sites) at the protein interface which mediate the interaction.

The identification of protein interaction sites is important for functional genomics, analysis of metabolic and signal transduction networks and also rational drug design. The first attempt to describe the characteristics of protein interaction sites was undertaken in 1975 by Chothia and Janin. With data from only three complexes, they suggested that the residues which form the interface are closely packed, tend to be hydrophobic and that complementarity may be an important factor in predicting which proteins can interact ***(10)***.

Later studies with larger datasets extended and developed their work to try to identify other characteristics of the interaction site that are sufficiently different from the rest of the protein to be identifiable, and thus be predictive. Further analysis of the hydrophobicity distribution of amino acids can be used to predict interaction sites since interacting regions tend to be the most hydrophobic clusters on the surface of the protein ***(11-14)***. This type of analysis yields a 60% success rate at predicting interacting sites. In general, hydrophobic residues such as Leu, Ile, Val, Phe, Tyr and Met are over-represented at interaction sites, whereas polar residues such as Lys, Asp and Glu (but not Arg) are under-represented ***(15,16)***. Other parameters which have been analyzed for their importance in identifying those residues in a protein which form the interaction site include the accessible surface area and residue composition ***(17,18)***. It has also become apparent that a distinction must be made between different types of complexes. Interaction sites on stable and transient complexes have different properties ***(18)***. A recent study ***(19)*** indicates that the the residue composition can be used to identify six different types of protein-protein interfaces, from domain-domain interfaces in the same protein to inter-protein contact surfaces.

Further studies, ***(18,20)*** using a six-parameter analysis (solvation potential, residue interface potential, hydrophobicity, planarity, protrusion and accessible surface area), have indicated that none of these parameters individually could be definitively used as a prediction method. Using a combined score from all six parameters yielded accurate predictions for 66% of 59 structures ***(21)***. All interfaces tend to be planar and be surface accessible, the other parameters differed between complex types. Computational resources for the prediction and analysis of protein-protein interactions in this way are described below (*see* **Methods 2.1.** and **Table 1**).

Shape complementarity is primarily used in docking studies which focus on finding the best fit of the two interacting proteins using rigid- and soft-body searches ***(22-24)***. Electrostatic complementarity between interfaces (**Fig. 1**) plays an important role in determining the best fit of two interacting proteins ***(23)***. Interfaces between antibody-antigen complexes and transient heterodimers tend to have the least shape complementarity, while homodimers, enzyme-inhibitor complexes and permanent heterodimers are the most complementary ***(18)***.

However, other research ***(25)*** has indicated that the chemico-physical properties of interacting surfaces are difficult to distinguish from those of the whole protein surface. Recently, it has been suggested that instead of using patch analysis, it may be better to use interface contacts ***(19)***, i.e., residues whose closest atoms are annotated in PDB as being less than 6Å apart. They argue that the analysis of surface patches may miss slightly buried residues with long side-chains, while other residues identified as being part of a patch may in fact not be important, or may not form contacts at all.

Other methods of predicting protein interaction sites include multiple sequence alignment and analysis of amino acid characteristics of neighboring residues using neural networks. Multiple sequence alignments can help identify specific family structures which are conserved within a subfamily but differ between subfamilies. These regions are interpreted as being interaction sites which may endow specificity of ligand interaction ***(26-28)***. Two groups ***(29,30)*** have trained neural networks with sequence profiles of spatial neighbors of a target residue with solvent exposure to predict whether a residue will be part of an interaction site. Both of their methods gave approximately a 70% accurate prediction rate. The validity of using sequence profiles has been verified by results which demonstrate that the majority of interacting residues are clustered in sequence segments of several contacting residues ***(31)***.

Recently, methods have been developed to validate predicted protein-protein interactions against experimentally determined 3D structures ***(32,33)***. Given a known three-dimensional structure, they map homologs of the interacting proteins onto the structure and using empirical potentials, test whether the homologous proteins preserve the interactions from the known structure. However, the number of experimentally determined structures for complexes is small, and of the 2,590 interactions predicted by large-scale methods, only 59 could be mapped onto their set of interacting complexes. Of these, 59% had domains that appeared to be in direct contact, thus increasing the probability that these predicted protein-protein interactions are biologically correct. Computational methods ***(34)*** for the prediction of protein-protein interactions based on this (and other structural approaches) are described below (*see* **Methods 2.1.**).

1.2. Genomic Context Approaches

1.2.1. Co-localization

One of the first methods for predicting protein-protein interactions from the genomic context of genes utilizes the idea of co-localization, or gene neighborhood (**Fig. 2A**). Such methods exploit the notion that genes which physically interact (or are functionally associated) will be kept in close physical proximity to each other on the genome ***(35-37)***. The most apparent case of this phenomenon involves bacterial and archaeal operons, where genes that work together are generally transcribed on the same polycistronic mRNA. In these cases, proteins involved in the same process or pathway are frequently encoded on the same polycistronic messenger.

Operons are rare in eukaryotic species ***(38,39)***. However, genes involved in the same biological process or pathway are frequently situated in close genomic proximity ***(36)***. It is hence possible to predict functional or physical interaction between genes that are repeatedly observed in close proximity (e.g. within 500 bp) across many genomes. This method has been successfully used to identify new members of metabolic pathways ***(36)***. Like many of the genome-context approaches, this method becomes more powerful with larger numbers of genomes. This approach and a number of online resources that implement it will be described in detail below (*see* **Methods 2.2.**).

1.2.2. Phylogenetic Profiles

A relatively simple, yet powerful, form of genomic context is the co-occurence of pairs of genes across multiple genomes. Two of the main driving forces in genome evolution are gene genesis and gene loss ***(40,41)***. The fact that a pair of genes remains together across many disparate species represents a concerted evolutionary effort that suggests that these genes are functionally associated (i.e. same biological process or pathway) or physically interacting. This criterion is less stringent than that of gene co-localization, where gene pairs must not only be present, but also situated close to each other on the genome. Homologous genes can be termed either *orthologs* or *paralogs*. In general the term ortholog is used to describe genes that are related by a speciation event, i.e. perform analogous functions in different organisms and are related to a single common ancestor gene in an ancestor species. The term paralog is used to describe homologous genes that have arisen following a gene duplication event, i.e., perform similar functions in the same organism. Classifying homologous genes as either paralogs or orthologs is difficult in the absence of accurate phylogenetic or speciation information ***(42)***. Classification of genes in this way allows the inference of a phylogenetic context for a given gene.

The analysis of phylogenetic context in this fashion has been termed *phylogenetic profiling* ***(43)***. These profiles can be as simple as a binary representation of the presence or absence of a gene across multiple genomes ***(43-45)*** (**Fig. 2B**). A library of these profiles may then be scanned to find genes that exhibit identical (or highly similar) phylogenetic patterns to each other. Pairs of genes detected in this fashion are hence candidates for physical interaction or functional association. This method has been used not only to infer physical interaction ***(43)***, but also to predict the cellular localization of gene products ***(46)***.

This system is not however without flaw. Firstly, the strength of any inference made using such profiles is heavily dependent on the number and distribution of genomes used to build the profile. A pair of genes with similar profiles across many of bacterial, archaeal and eukaryotic genomes are much more likely to interact than genes found to co-occur in a small number of closely related species. Secondly, evolutionary processes such as lineage-specific gene loss, horizontal gene transfer, non-orthologous gene-displacement ***(47)*** and the extensive expansion of many eukaryotic gene families can make orthology assignment across genomes very difficult. However, given the increasing number of completely sequenced genomes, the accuracy of these predictions is expected to improve over time. The details of this approach and online-resources for phylogenetic profile-based prediction of protein interaction are described below (*see* **Methods 2.3.**).

1.2.3. Gene Fusion

Genome context approaches to the prediction of protein-protein interaction also include the analysis of gene fusion across complete genomes. This method is complementary to both co-localization of genes and phylogenetic profiles and uses both gene location and phylogenetic analysis to infer function or interaction. A gene fusion event represents the physical fusion of two separate parent genes into a single multi-functional gene. This is the ultimate form of gene co-localization, i.e., interacting genes are not just kept in close proximity on the genome, but are physically joined into a single entity (**Fig. 2A**). It has been suggested that the driving force behind these events is to lower the regulational load of multiple interacting gene products ***(48)***. Gene fusion events hence provide an elegant way to computationally detect functional and physical interactions between proteins ***(48,49)***.

Gene fusion events are detected by cross-species sequence comparison. Fused (composite) proteins in a given reference genome are detected by searching for un-fused component protein sequences, that are homologous to the reference protein, but not to each other. These un-fused query sequences align to different regions of the reference protein, indicating that it is a composite protein resulting from a gene fusion event ***(48)***. Once again, predictions of this type are complicated by a number of issues. The largest hindrance is the presence of so called *promiscuous* domains. These domains (such as helix-turn-helix (HTH) and DnaJ) are highly abundant in eukaryotic organisms. The domain complexity of eukaryotic proteins coupled with the presence of promiscuous domains and large degrees of paralogy can hamper the accurate detection of gene fusion events ***(50)***.

Although the method is not generally applicable to all genes, i.e., it requires that an observable fusion event can be detected between gene pairs, it has been successfully applied to a large number of genomes (including eukaryotes) ***(51)***. The basic gene fusion detection method and online resources such as the **AllFUSE** database ***(51)***, will be described in detail below (*see* **Methods 2.4.**).

1.2.4. In-silico two-hybrid

The *in-silico two-hybrid* (**i2h**) approach has much in common with the other genome-context approaches, but also indirectly assesses structural properties of proteins that potentially interact. It has previously been shown that a mutation in the sequence of one protein in a pair of interacting proteins is frequently mirrored by a compensatory mutation in its interacting partner (**Fig. 2C**). The detection of such correlated mutations can not only be used to predict protein-protein interactions, but also has the potential to identify specific residues involved at the interaction sites ***(52)***.

Previous analyses ***(53)*** involved searching for correlation of residue mutations between sequences in the same protein family alignment (intra-family). The *in-silico* two-hybrid method extends this approach by searching for such mutations across different protein families (interfamily). Prediction of protein-protein interactions using this approach is achieved by taking pairs of protein family alignments and concatenating these alignments into a single cross-family alignment. A position-specific matrix is then built from this alignment, and a correlation function is then applied to detect residues which are correlated both within and across families. Correlated sites that potentially indicate protein interaction are returned with a score. The method suffers due to the computational complexity of constructing the large numbers of alignments needed, and poor quality alignments can dramatically increase noise in the procedure ***(52)***. However the method is similar to the gene fusion approach, as a single accurate prediction between two proteins can infer interaction between all members of both families used. This approach is not discussed in the Methods section, as currently the method is not freely available (*see* **Note 1**).

1.3. Biological Context Approaches

High throughput experimental techniques now provide access to a more detailed view of biological processes at a genomic level. Gene expression analysis allows one to not only determine which genes are active in a given state, but also sets of genes which are co-regulated in many different states. It has been shown that many interacting proteins are co-expressed according to microarray analyses ***(4-6)***. Current gene-expression methods now allow for every coding gene of a genome to be placed on a single microarray, allowing the activity of every gene to be monitored across different states or time-points. Although these methods cannot directly be used to determine whether or not two proteins interact or not, a number of computational approaches have been developed that use this information towards the prediction of protein-protein interaction and gene regulatory networks ***(4-6)***. Other high-throughput experimental techniques such as yeast two-hybrid specifically test a bait protein for interactions against a set of prey proteins. The bait and prey consist of fusion constructs that activate a reporter gene if they interact with each other. While this method is not as accurate as other techniques such as co-immunoprecipitation, affinity chromatography or gel-overlay assays, it can be applied rapidly to genome-scale studies of protein-protein interactions.

Many of these high-throughput methods for investigating the biological context of genes and proteins are inherently noisy. For example, some proteins in yeast two-hybrid assays appear to detect a large number of spurious interactions (false-positives). Gene expression techniques suffer from a number of problems also, such as cross hybridization and poor signal-to-noise ratios. Recently however, research has shown that multiple datasets pertaining to the biological context of genes and proteins can be combined using machine learning techniques ***(8)***. Using Bayesian network analysis it is hence possible to computationally combine multiple noisy datasets in such a way that protein-protein interactions can be more reliably predicted. In this method each source of interaction evidence is compared against samples of known positive (proteins in the same complex) and negative (proteins in different cellular locations) interactions, allowing a statistical reliability index to be built for each data source. When this information is applied genome-wide, a prediction can be made for every protein pair in a genome by combining different sets of independent evidence according to their calculated reliability. Protein interactions predicted in this way have been shown to be as reliable as pure experimental techniques, while simultaneously covering a larger proportion of genes than most experimental methods ***(8)***.

A number of available resources for protein-protein interaction data, gene expression data and Bayesian network analysis of multiple interaction datasets will be described below (*see* **Methods 2.5.**).

1.4. Data-sources and Visualization techniques

Computational biology is a data-rich research field. The advent of complete genome sequencing and high-throughput experimental techniques has created an enormous amount of data. In order for these data to be both informative and useful, they must be stored in a sensible and accessible way, and tools must be made available to visualize and exchange this information. A number of initiatives are tackling these problems by creating freely accessible databases storing a wide variety of biological information including protein-protein interactions. Recently, a number of research groups have created visualization tools for biological networks. These tools provide a new way to analyze protein-protein interaction networks, provide a multitude of different ways to represent interactions and can overlay other biological information onto these networks. A number of databases that store protein-protein interactions, molecular complexes and pathways will be described later (*see* **Methods 2.6.** and **Table 1**). Finally, we will detail methods for the visualization and analysis of protein-protein interaction networks. (*see* **Methods 2.7.**).

2. Methods

In this section we will describe computational resources and methods for the prediction of protein-protein interactions. These methods will be detailed in chronological order. Within each section a number of on-line computational resources are described that allow one to perform this type of analysis interactively. Where possible (mostly for genomic context based approaches), detailed computational protocols will also be provided. Resources mentioned in this section are further summarized below (*see* **Table 1)**.

2.1. Structure based prediction of interactions

The *Protein-Protein Interaction Server* at University College London (UCL) provides a simple web-based interface for exploring protein-protein interaction interfaces, given three-dimensional structures ***(18)***. This server takes into account the following information for interaction analysis: accessible surface area, planarity, length & breadth, secondary structure, hydrogen bonds, salt bridges, gap volume, gap volume index, bridging water molecules and interface residues. This resource (**Table 1**) is very useful for exploring the protein-protein interaction potential of two protein structures identified through docking or shape-complementarity.

The structural bioinformatics group at EMBL Heidelberg provides the **InterPreTS** server for protein-protein interaction prediction ***(33)***. Using this resource (**Table 1**) one can submit pairs of sequences that are then compared to the three-dimensional structures of known protein-protein interactions. This resource utilizes a pre-built *Database of Interacting Domains* (**DBID**) and an empirical scoring system to test whether a sequence pair fits a known three-dimensional structure of an interacting pair of proteins.

2.2. Gene-neighborhood based interaction prediction

Co-localization of genes across multiple genomes provides a fingerprint that they may physically interact ***(36)***. Analysis of conserved gene locations across multiple genomes (**Fig. 2A**) can hence be used to predict protein interaction networks and metabolic pathways ***(54)***. A number of excellent resources exist that allow one to determine whether two proteins may interact using this approach. The most notable of these are **STRING** (Search Tool for Recurring Instances of the Neighborhood of Genes) ***(55)*** and **WIT** (What is There?) ***(56)***. The **STRING** database (**Fig. 3**) provides a web interface giving comprehensive access to gene neighborhood information ***(57)*** for 356,775 genes in 110 complete genomes (**Fig. 3**). Similarly, the **Predictome** database at Boston University ***(58)*** provides a comprehensive web interface to predictions of this type. The **WIT** database provides access to protein family information, metabolic pathway reconstruction and gene co-localization information. Using these resources allows detailed pre-computed gene neighborhood information to be analyzed for evidence of protein-protein interaction

(**Table 1**). The actual protocols used for these analyses can vary considerably, a general protocol adapted from WIT ***(56)*** is described below:

1. In order to assess whether pairs of orthologous genes share a common gene neighborhood across multiple genomes one needs a) protein sequences / genomic locations and b) orthology mappings between proteins from multiple genomes.

2. Orthology mappings are generated by searching for pairs of *close bi-directional best hits* (PCBBH). These are a specific form of bi-directional best hit (*see* **Methods 2.3.**), a commonly used method for orthology assignment. For a given pair of proteins □ and □ in genome X, a bi-directional best hit to genes □' and □' in genome Y is defined as follows:

    a. The best BLAST hit for protein □ in genome X is protein □' in genome Y.
    b. The best BLAST hit for protein □ in genome X is protein □' in genome Y.
    c. The genes of proteins □ and □ are situated within 300bp in genome X.
    d. The genes of proteins □' and □' are situated within 300bp in genome Y.

3. Genes that satisfy the above criteria can be considered as having a conserved gene neighborhood across two genomes. When this procedure is repeated across multiple genomes it becomes possible to identify genes which are significantly co-localized across many genomes, and are hence likely to either physically interact or be functionally associated.

4. The PCBBH criteria are quite strict, and it is also possible to perform the procedure using *Pairs of Close Homologs* (PCHs) (*see* **Note 2**).

5. Sets of PCBBHs or PCHs in multiple genomes are typically scored for significance based on the number and phylogenetic distribution of genomes in which they are co-localized. Phylogenetic distance can be estimated by examining a 16S rRNA phylogenetic tree.

6. A common score (coupling score) for the likelihood that two genes interact based on summing individual scores from multiple genomes is then calculated.

7. Finally, candidate genes that have significant coupling scores are candidates for either physical interaction, or functional association.

2.3. Phylogenetic profile based prediction of interaction

Phylogentic profile based prediction of protein interactions (**Fig. 2B**) has been shown to be an accurate and widely applicable method. Perhaps the easiest way to utilize this information for prediction of protein interaction is to use precomputed phylogenetic profiles for proteins of interest. The *Clusters of Orthologous Groups* (**COGs**) resource at the National Center for Biotechnology Information (NCBI) contains large numbers of profiles for a variety of bacterial and archaeal organisms and also *S. cerevisiae* ***(59,60)***. Other excellent resources for combined computational predictions of protein interactions using phylogenetic profiles are available from the **STRING** ***(55)*** resource at EMBL Heidelberg and from **Predictome** ***(58)*** (**Table 1**). Using the web interfaces to these resources, it is relatively straightforward to find groups of proteins with similar or identical phylogenetic profiles, indicating proteins that physically interact or are functionally associated (**Fig. 3**). For a more detailed analysis of specific proteins of interest a general protocol is described below:

1. For each genome to be analyzed, a FASTA sequence file containing all protein sequences is assembled.

2. All protein sequences in each genome are compared against all other sequences using a sequence similarity search algorithm such as BLASTp ***(61)***. A variety of other sequence similarity seach tools could also be used at this step (*see* **Note 3**).

3. Orthology between proteins in different genomes is assigned as follows:

    - Two proteins (from different genomes) are orthologous if they were each other's highest scoring BLAST hit when searched against the other genome. This is frequently referred to as a *bidirectional best hit* (BBH).

    - This process is repeated to assign (if possible) an ortholog for each protein in a given genome, to a protein in all other genomes.

4. All orthology assignments made in this way are stored for post-processing.

5. A phylogenetic profile for a protein can then be constructed by representing the presence or absence of an ortholog for that protein across all genomes analyzed. Frequently, this is represented by a simple binary vector with '1' indicating presence and '0' representing absence of a gene in each genome (**Fig. 2B**) (*see* **Note 4**).

6. All profiles are compared to all other profiles using a clustering procedure. A distance measure (such as Pearson correlation of Euclidean distance) between each profile and all other profiles is used to group profiles according to how similar they are (*see* **Note 5**).

Finally, protein profiles that are highly similar or identical to each other represent candidate proteins that physically or functionally interact.

2.4. Gene fusion prediction of protein interactions

Gene fusion is a relatively common evolutionary phenomenon ***(51)***. A detected gene fusion between two genes indicates that their protein products may physically interact or be involved in the same biological process or pathway ***(48,49)***. One extreme example of this is the aromatic amino acid biosynthesis pathway in *S. cerevisiae*. In yeast a single fused gene encodes the entire pathway of these five normally separate genes ***(48)***. Prediction of protein interactions using gene fusion has been successful in a number of areas, including the prediction of novel protein interactions involved in important biological processes in *Drosophila melanogaster* ***(62)***.

A comprehensive set of fused genes and inferred protein-protein interactions is available from the **AllFUSE** database ***(51)*** at the European Bioinformatics Institute (EBI), the **STRING** database at EMBL Heidelberg ***(55)*** (**Fig. 3**) and the **Predictome** database at Boston University ***(58)*** (**Table 1**). Using the **AllFUSE** resource one can search for potential interactions for a given protein sequence from a database of 24 complete genomes. A general protocol for gene fusion based prediction of protein-protein interactions can be described as follows ***(48)***:

1. This analysis requires two genomes, a *query* and a *reference*. One searches for gene fusion (composite) proteins in the reference genome using protein sequences from the query genome. Sequences from both genomes need to be assembled into FASTA format for this analysis.

2. Each protein in the query genome is then interactively searched against each protein from the reference genome using a sequence similarity search tool such as BLASTp ***(61)*** using an expectation-value (E-value) threshold to eliminate similarities which may have arisen by chance.

3. All significant similarities detected in this way are then stored in a binary matrix which for each protein pair stores '1' for significant similarity or '0' for no detectable similarity. The matrix may be symmetrified by post-processing with a more sensitive sequence search tool such as Smith-Waterman ***(63)*** to clear up ambiguities.

4. Finding evidence of a gene fusion event in the reference species extends of the previous symmetrification problem to one of transitivity ***(48)***. In this case one searches for instances where query proteins **A** and **B** match a reference protein **C**, but do not match each other (i.e. **A**□ **C**; **B**□ **C** but **A**≠**B**). These triangular inequalities are resolved once again by using the more accurate Smith-Waterman algorithm to double check that no detectable significant similarity exists between **A** and **B**. Further analysis using alignment geometry can then verify that proteins **A** and **B** are orthologous to different regions of a composite fusion protein but not to each other ***(64)***.

5. Candidate fusion proteins detected in this way provide evidence that proteins **A** and **B** may physically interact.

Although this method is not generally applicable to all genes, and suffers from the high levels of paralogy usually present in eukaryotic genomes ***(65)***. This approach has been shown to have an accuracy as high as 90% and readily detects well known interacting proteins (e.g. tryptophan synthase □ and □ subunits) and many proteins previously shown to form complexes. As such this method represents a useful way to build interaction networks for proteins of interest within and across genomes.

2.5. Prediction of protein interactions from high-throughput biological datasets

Gene expression analysis allows for all genes from a given genome to be placed on a single microarray, allowing many gene-expression experiments to be carried out rapidly and in parallel. Recently, efforts have been made to standardize data formats for reporting the results of gene expression experiments. The *Minimum Information About a Microarray Experiment* (**MIAME**) ***(66)*** standard allows different laboratories to effectively and accurately exchange microarray expression information. Using such standards, it has become easier for a number of publicly accessible resources to distribute microarray data (**Table 1**).

The *Stanford Microarray Database* (**SMD**) ***(67)*** provides access to raw data from public microarray experiments, as well as a number of software tools for utilizing this data. Currently, 140 experiments are indexed in the **SMD** web resource. The MicroArray group at the European Bioinformatics Institute provides **ArrayExpress** ***(68)***, a publicly available gene expression data in **MIAME** format for over 66 publicly available experiments and also integrated tools for expression profile analysis. Finally, the *Gene Expression Omnibus* (**GEO**) ***(69)*** database at the NCBI contains data from over 300 large-scale publicly available microarray and SAGE experiments, for which all data is linked into the NCBI protein, nucleotide and genomic databases.

Using these resources, it is hence possible select a number of datasets for an organism of interest, and extract gene expression profiles for some or all genes. Proteins whose genes exhibit very similar patterns of expression across multiple states or experiments ***(70)*** may then be considered candidates for functional association and possibly direct physical interaction ***(4-6)***. Gene expression analysis becomes much more reliable with more expression data. For example, genes that have high correlation across 10 experiments are much more likely to be related functionally than genes correlating across two experiments. Gene expression data is relatively susceptible to noise, and great care must be taken to minimize and filter this from any analysis. This data can, however, be very powerful when combined with analyses involving regulatory network reconstruction, and with other methods of detection of functional association and interaction of proteins ***(8)***.

The Bayesian networks approach (*see* **Introduction 1.3.**), which combines data from multiple biological datasets is a useful way to minimize this noise and perform reliable protein-protein interaction prediction in *S. cerevisiae* ***(8)***. Validation of the method indicates that it can successfully recover large numbers of previously known protein-protein interactions (**Fig. 4**) and many novel interaction predictions. The results of this analysis are available from the **GeneCensus** web site at Yale University (**Table 1**). These predictions are remarkable as they illustrate that combining multiple independent and noisy datasets in an intelligent way does not necessarily increase noise in the combined protein interaction predictions (assuming orthogonal error between datasets). This is also an excellent example of a combined computational and experimental approach, as interactions predicted using this approach appear to be more reliable than many pure experimental approaches ***(8)***.

2.6. Tools for Protein-Protein Interaction Visualization

Network and pathway visualization tools are computer programs that can automatically generate a diagram of a network or pathway. Perhaps the simplest such representation of a protein-protein interaction network is a graph composed of nodes (proteins) connected by edges (interactions). Some of the first visualization tools were developed for browsing metabolic pathways. For example, a pathway drawing tool is present in the **ACeDB** database ***(71)*** and in **EcoCyc** ***(34)***. In many cases these representations are *clickable* so that one can select a member of a pathway or a small molecule and get further information about that entity. Many of these initial visualization tools are static, and generated semi-automatically. The *Kyoto Encyclopedia of Genes and Genomes* (**KEGG**) ***(72)***, **BioCarta** and **SigPath** ***(73)*** websites (**Table 1**), are examples of this type of visualization. Other more advanced methods can dynamically generate pathway diagrams from raw information in a biological database, such as the **EcoCyc** and **WIT** databases (*see* **Table 1**).

Recently, a number of purely automatic and general algorithms have been developed for visualizing biological networks. These tools rely on a layout algorithm to organize a graph of nodes and edges into an *aesthetically pleasing* layout. In graph terms this usually means minimizing the number of edges that cross each other, and grouping groups of nodes that are highly connected to each other. Typically, a well-organized graph layout will allow the user to identify global features of their data that may not have been previously apparent. An example layout algorithm is the *Spring Embedder* algorithm. This method models the graph as a physical system where nodes are spheres connected by springs (edges). Nodes are initially organized in a random state, and forces between connected spheres (due to springs), push the system into a lower-energy more stable state. Other methods such as the *Weighted Fruchterman-Rheingold* algorithm ***(64)***, represent the graph as a system of nodes which exert an attractive force (similar to a spring) between nodes connected by an edge and a distance-dependant repulsive force between all nodes. Additionally the weighted algorithm allows the attractive forces between node to be modulated using weights, and the energy of the entire system is controlled using a temperature function. Other layout algorithms, can involve arranging nodes hierarchically, in a circular fashion or in less structured formats. It is important to choose the best layout algorithm for the type of graph being visualized. For example, a highly connected interaction network will not assume a meaningful graph layout when a hierarchical layout algorithm is used.

Two of the most commonly used visualization tools for biological networks are **BioLayout** ***(74)*** and **Cytoscape** ***(75)*** (**Table 1**). Both of these tools are written using the JAVA (*see* **Note 6**) programming language and are hence portable across a wide variety of computer environments. Both tools also allow the interactive editing of graphs, through the movement of nodes, node labeling and the ability to change the appearance of nodes and edges. Additionally both tools can export publication quality high-resolution graph images (*see* **Note 7**). **BioLayout** utilizes the weighted Fruchterman-Rheingold layout algorithm, and has a number of options for graph customization, data-overlay, export and graph analysis (**Fig. 5**). **Cytoscape** provides a number of different layout algorithms for producing useful visualizations and a number of *plugins* and import options for representing data such as gene expression (**Fig. 6**). Specifically, circular, hierarchical, organic, embedded and random layouts are available. Circular and hierarchical algorithms try to lay out a network as their names suggest. Organic and embedded are two versions of a force-directed layout algorithm. Types of plugins that are currently available for **Cytoscape** include one that allows reading **PSI** files (*see* **Methods 2.7.**) and one called **ActiveModules** that finds regions of a molecular interaction network that are correlated across multiple gene expression experiments. Both of these methods are suitable for small to medium sized networks (less than 1000 nodes), although it may not be long before both layout and visualization techniques become available for the analysis of much larger graphs.

2.7. Data resources for protein-protein interactions

Current computational and experimental methods for protein-protein interaction prediction have been generating large amounts of data. It is imperative that this data be stored in a consistent and reliable way so that it may be useful for biological research. A number of databases are now publicly available for making this information accessible. Two of the largest and most comprehensive interaction databases now available are the *Biomolecular Interaction Network Database* (**BIND**) ***(76)*** and the *Database of Interacting Proteins* (**DIP**) ***(77)***. **DIP** is based at UCLA and currently contains over 18,000 experimentally determined protein-protein interactions (mostly from high-throughput *S. cerevisiae* experiments) for over 7,000 proteins in 104 organisms. Interactions in **DIP** are curated both manually (by expert curators) and automatically (text-mining approaches). **BIND**, at the University of Toronto, not only stores and curates pair-wise protein-protein interactions, but also molecular complex information and biological pathways. Currently, **BIND** contains over 21,349 protein-protein and protein-DNA interactions, 1,334 molecular complexes and 8 pathways encompassing 28 genomes and over 6,000 proteins.

A number of initiatives are currently underway to ensure that these data from different interaction databases are stored in a consistent and exchangeable format. The *Proteomics Standards Initiative* (**PSI**) ***(78)*** has created a standard format for the exchange of protein-protein interaction data, while the **BioPAX** format aims to capture protein-protein interactions, molecular complexes and pathway information in a single consistent ontology and exchange format. Access information for **DIP**, **BIND** and a number of other interaction databases are detailed further below (*see* **Table 1**).

3. Notes

1. Unfortunately, the online web-server for *in-silico two-hybrid* predictions at **http://www.pdg.cnb.uam.es/i2h/** is not presently available, although the **Plotcorr** program for analysis of correlated mutations is available at: **http://www.pdg.cnb.uam.es/pazos/plotcorr.html**
2. *Pairs of close homologs* (PCHs) can be defined as follows: (a) A significant BLAST hit exists for protein □ in genome X and protein □’ in genome Y. (b) A significant BLAST hit exists for protein □ in genome X and protein □’ in genome Y. (c) The genes of proteins □ and □ are situated within 300bp in genome X. (d) The genes of proteins □’ and □’ are situated within 300bp in genome Y.
3. While BLAST is useful for many analyses of this type, it is also possible to use more sensitive algorithms such as PSI-BLAST, HMMER or Smith-Waterman. Although these methods tend to be far more computationally intensive, the may produce more accurate predictions.
4. Phylogenetic profiles do not necessarily have to be binary representations. It would also be possible to generate profiles that express a score or expectation value that a homolog is present in a given genome instead of simply ‘1’ and ‘0’.
5. This type of analysis is very easy to perform using common mathematical analysis tools or the **PEARSON** function from Microsoft Excel™.
6. The Java™ environment is commonly preinstalled on many computer systems. If not already installed, it can be obtained at: **http://java.sun.com/**
7. Graph images of protein-protein interaction networks can be exported in a number of ways. Capturing screenshots of either application will most likely result in a poor quality low-resolution image. For publication quality images, it is generally best to export images as a vector graphics format such as PDF.
8. The PyMOL molecular graphics package is freely available for a variety of platforms at: **http://pymol.sourceforge.net/**

Acknowledgements

The authors would like to thank Ronald Jansen for providing information about Bayesian network based prediction of protein-protein interactions and the graph used for Figure 4.

Resource	Type of Resource	WWW Address (URL)	Ref
Structural Context Interaction Prediction			
Protein-Protein Interaction Server	Structure based interaction prediction	http://www.biochem.ucl.ac.uk/bsm/PP/server/	(18)
InterPreTS	Structure based interaction prediction	http://www.russell.embl.de/interprets/	*(33)*
Genomic Context Interaction Prediction			
AllFUSE	Gene fusions	http://www.ebi.ac.uk/research/cgg/allfuse/	*(51)*
STRING	Gene Co-Localization, gene-fusion, phylogenetic profiles	http://www.bork.embl-heidelberg.de/STRING/	*(55)*
WIT	Orthology/phylogenetic profiles / gene co-localization	http://wit.mcs.anl.gov/WIT2/	*(56)*
Predictome	Gene Co-Localization, gene-fusion, phylogenetic profiles	http://predictome.bu.edu/	*(58)*
COGs	Orthology/phylogenetic profiles	http://www.ncbi.nlm.nih.gov/COG/	*(59)*
Biological Context Interaction Prediction			
GeneCensus	Combined predictions (bayesian network)	http://genecensus.org/intint/	*(8)*
Pathway Databases			
EcoCyc	Metabolic pathway analysis	http://ecocyc.pangeasystems.com/ecocyc/	*(34)*
KEGG	Metabolic / regulatory pathway analysis and reconstruction	http://www.genome.ad.jp/kegg/	*(72)*
SigPath	Signalling pathways	http://www.sigpath.org/	*(73)*
MIPS	Pathways, complexes, cellular locations	http://www.mips.biochem.mpg.de/proj/yeast/pathways/index.html	*(79)*
Protein Interaction Databases			
BIND	Interactions, complexes, pathways	http://www.bind.ca/	*(76)*
DIP	Database of protein interactions	http://dip.doe-mbi.ucla.edu/	*(77)*
INTACT	Database of protein interactions	http://www.ebi.ac.uk/intact/index.html	*(80)*
MINT	Database of protein interactions	http://160.80.34.4/mint/	*(81)*
Gene-Expression Databases			
SMD	Gene expression data	http://genome-www5.stanford.edu/	*(67)*
Array Express	Gene expression data	http://www.ebi.ac.uk/arrayexpress/	*(68)*
GEO	Gene expression data	http://www.ncbi.nlm.nih.gov/geo/	*(69)*
Visualization Tools for Protein Interactions			
BioLayout	Interaction Network Visualization	http://www.biolayout.org/	*(74)*
Cytoscape	Interaction Network Visualization	http://www.cytoscape.org/	*(75)*

Table 1: Methods and databases for computational prediction of protein-protein interactions.

Figure Legends

Figure 1: Three-dimensional structure of the T7 bacteriophage RNA polymerase complexed with T7 lysozyme. The multi-colored structure on the left is RNA polymerase, shown with a transparent blue molecular surface. The lysozyme is shown on the right in grey with its associated transparent surface. The interaction interface is highlighted in yellow on both surfaces. This figure was produced using PDB structure 1ARO and PyMol (*see* **Note 8**).

Figure 2: Overview of genome context approaches. A) Gene neighborhood plots for eight complete genomes, showing a pair of genes (red and blue) which are in close physical proximity in all eight genomes. A gene fusion event between two genes (yellow and light blue) in two genomes is also shown. B) Example phylogenetic profiles of selected genes from the previous panel. These three pairs of genes have the same patterns of co-occurrence in all eight genomes, and may physically interact based on this evidence. C) Two protein family alignments are shown with conserved regions highlighted (in red and blue). Correlated mutations (shown in green) are present in two identical sub-trees for each family, which indicates that these sites may be involved in mediating interactions between proteins from each family.

Figure 3: Screenshots from the STRING web resource. The left panel illustrates the STRING representation of gene neighborhood and gene fusion. The right panel shows a typical phylogenetic profile for multiple genes and genomes. Finally, the inset shows a predicted protein interaction map generated from gene neighborhood, gene fusion and phylogenetic profile methods combined.

Figure 4: Bayesian network predictions of protein-protein interactions. Experimentally validated *gold-standard* protein-protein interactions (blue and green lines) between *S. cerevisiae* proteins (green dots) are shown as an interaction network. Bayesian network analysis prediction of protein-protein interaction successfully recovers a significant subset of these interactions (green lines). The gold-standard interactions are derived from MIPS and well-known complexes are annotated.

Figure 5: Example graph from BioLayout. This graph illustrates a genetic regulatory network of *E. coli* genes. Genes are represented by circles (nodes) connected by regulatory interactions represented by lines (edges). Nodes are colored according to biochemical pathway assignments. Nodes in the center of the graph are not labeled for clarity.

Figure 6: Example graph from Cytoscape. A number of the important features of CytoScape are represented in this graph layout. Nodes in this case represent genes and edges represent either genetic (green, cyan) interactions, protein-protein interactions (blue) or protein-DNA interactions (red). Nodes are colored according to the gene expression of that gene in a Gal4 knockout experiment, with blue representing highly significant fold-change of a gene, and red indicating no significant fold-change. Node shapes are determined by the annotation of each gene, diamonds for signal transduction genes, triangles for meiosis, Pol III transcription, mating response and DNA repair. Circles represent genes that were not assigned to any of these categories.

References

1. Mendelsohn, A. R., and Brent, R. (1999) Protein interaction methods - toward an endgame. *Science* **284**, 1948-1950.
2. Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000) Protein function in the post-genomic era. *Nature* **405**, 823-826.
3. Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000) Exploitation of gene context. *Curr Opin Struct Biol* **10**, 366-370.
4. 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.
5. Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-486.
6. Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46.
7. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86.
8. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453.
9. Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., and Abola, E. E. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* **54**, 1078-1084.
10. Chothia, C., and Janin, J. (1975) Principles of protein-protein recognition. *Nature* **256**, 705-708.
11. Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. (2000) A fast method to predict protein interaction sites from sequences. *Journal of Molecular Biology* **302**, 917-926.
12. Korn, A. P., and Burnett, R. M. (1991) Distribution and Complementarity of Hydropathy in Multisubunit Proteins. *Proteins-Structure Function and Genetics* **9**, 37-55.
13. Young, L., Jernigan, R. L., and Covell, D. G. (1994) A Role for Surface Hydrophobicity in Protein-Protein Recognition. *Protein Science* **3**, 717-729.
14. Mueller, T. D., and Feigon, J. (2002) Solution structures of UBA domains reveal a conserved hydrophobic surface for protein-protein interactions. *Journal of Molecular Biology* **319**, 1243-1255.
15. Lijnzaad, P., and Argos, P. (1997) Hydrophobic patches on protein subunit interfaces: Characteristics and prediction. *Proteins-Structure Function and Genetics* **28**, 333-343.
16. Janin, J., Miller, S., and Chothia, C. (1988) Surface, Subunit Interfaces and Interior of Oligomeric Proteins. *Journal of Molecular Biology* **204**, 155-164.
17. Argos, P. (1988) An Investigation of Protein Subunit and Domain Interfaces. *Protein Engineering* **2**, 101-113.

18. Jones, S., and Thornton, J. M. (1996) Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13-20.
19. Ofran, Y., and Rost, B. (2003) Analysing six types of protein-protein interfaces. *Journal of Molecular Biology* **325**, 377-387.
20. Jones, S., and Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology* **272**, 121-132.
21. Jones, S., and Thornton, J. M. (1997) Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology* **272**, 133-143.
22. Lawrence, M. C., and Colman, P. M. (1993) Shape Complementarity at Protein-Protein Interfaces. *Journal of Molecular Biology* **234**, 946-950.
23. Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology* **272**, 106-120.
24. Shoichet, B. K., and Kuntz, I. D. (1991) Protein Docking and Complementarity. *Journal of Molecular Biology* **221**, 327-346.
25. Aloy, P., and Russell, R. B. (2002) Potential artefacts in protein-interaction networks. *Febs Letters* **530**, 253-254.
26. Casari, G., Sander, C., and Valencia, A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol* **2**, 171-178.
27. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**, 342-358.
28. Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol* **271**, 511-523.
29. Zhou, H. X., and Shan, Y. B. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins-Structure Function and Genetics* **44**, 336-343.
30. Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry* **269**, 1356-1361.
31. Ofran, Y., and Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *Febs Letters* **544**, 236-239.
32. Aloy, P., and Russell, R. B. (2002) Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5896-5901.
33. Aloy, P., and Russell, R. B. (2003) InterPreTS: protein Interaction Prediction through Tertiary Structure. *Bioinformatics* **19**, 161-162.
34. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28**, 56-59.
35. Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* **44**, 66-73.
36. Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-328.

37. 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.
38. Zorio, D. A., Cheng, N. N., Blumenthal, T., and Spieth, J. (1994) Operons as a common form of chromosomal organization in C. elegans. *Nature* **372**, 270-272.
39. Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* **20**, 480-487.
40. Snel, B., Bork, P., and Huynen, M. A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**, 17-25.
41. Kunin, V., Cases, I., Enright, A. J., de Lorenzo, V., and Ouzounis, C. A. (2003) Myriads of protein families, and still counting. *Genome Biol* **4**, 401.
42. Ouzounis, C. (1999) Orthology: another terminology muddle. *Trends Genet* **15**, 445.
43. 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.
44. Ouzounis, C., and Kyrpides, N. (1996) The emergence of major cellular processes in evolution. *FEBS Lett* **390**, 119-123.
45. Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95**, 6239-6244.
46. Marcotte, E. M., Xenarios, I., van Der Bliek, A. M., and Eisenberg, D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A* **97**, 12115-12120.
47. Galperin, M. Y., and Koonin, E. V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* **18**, 609-613.
48. Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.
49. 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.
50. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584.
51. Enright, A. J., and Ouzounis, C. A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol* **2**, RESEARCH0034.
52. Pazos, F., and Valencia, A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219-227.
53. Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
54. Snel, B., Bork, P., and Huynen, M. A. (2002) The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A* **99**, 5890-5895.
55. Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**, 3442-3444.

56. Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E., Jr., 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.
57. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258-261.
58. Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseris, J., and DeLisi, C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res* **30**, 306-309.
59. Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* **278**, 631-637.
60. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.
61. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.
62. Iliopoulos, I., Enright, A. J., Poullet, P., and Ouzounis, C. (2003) Mapping functional associations in the entire genome of Drosophila melanogaster. *Comparative and Functional Genomics* **4**, 337-341.
63. Smith, T. F., and Waterman, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197.
64. Enright, A. J. (2002) Computational Analysis of Protein Function in Complete Genomes. *Ph.D. Thesis*, University of Cambridge pp. 241.
65. Enright, A. J., Kunin, V., and Ouzounis, C. A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31**, 4632-4638.
66. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-371.
67. Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C., Schroeder, M., Brown, P. O., Botstein, D., and Sherlock, G. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res* **31**, 94-96.
68. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S. A. (2003) ArrayExpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **31**, 68-71.
69. Edgar, R., Domrachev, M., and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210.

70. 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 U S A* **95**, 14863-14868.
71. Walsh, S., Anderson, M., and Cartinhour, S. W. (1998) ACEDB: a database for genome information. *Methods Biochem Anal* **39**, 299-318.
72. Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30.
73. Campagne, F., Neves, S., Chang, C., Skrabanek, L., Ram, P. T., Iyengar, R., and Weinstein, H. (2003), In Press.
74. Enright, A. J., and Ouzounis, C. A. (2001) BioLayout - an automatic graph layout algorithm for similarity visualization. *Bioinformatics* **17**, 853-854.
75. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*, In Press.
76. Bader, G. D., Betel, D., and Hogue, C. W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-250.
77. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-291.
78. Orchard, S., Hermjakob, H., and Apweiler, R. (2003) The proteomics standards initiative. *Proteomics* **3**, 1374-1376.
79. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-34.
80. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and R., A. (2004) IntAct - an open source molecular interaction database. *Nucleic Acids Res*, In Press.
81. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett* **513**, 135-140.