Defining interacting partners for drug discovery

Matteo Pellegrini
Protein Pathways Inc., USA

Over the past few years several technologies have been developed to determine interacting partners of proteins. The techniques fall into two broad categories: direct and indirect. Experimental techniques have been developed to directly probe protein interactions by monitoring protein binding events. These techniques include the two-hybrid approach, protein fragment complementation assays, co-purification techniques and protein chips. In addition to these methodologies, several approaches have also emerged over the past few years to deduce indirect couplings between proteins. These couplings do not necessarily imply that two proteins are bound within the cell however they do provide evidence that perturbing one protein is likely to significantly perturb the function of its partner. These couplings may be deduced by studying the evolution of protein pairs, estimating the degree of correlated transcription of two genes or by searching for synthetic lethal pairs. In all cases, protein interactions and protein couplings are being used to advance drug discovery by providing detailed information on protein functions and by suggesting novel targets that act within biochemical pathways implicated in disease.

Keywords: protein interactions, two hybrid, TAP tags, phylogenetic profiles, co-expression, synthetic lethal mutations

1 Introduction

1.1 Using Protein Interactions to Guide Target Selection

Understanding the interactions between molecules within cells should dramatically improve our ability to design new drugs. This assumption is based on the belief that most molecules within in a cell participate in multiple interactions that form vast cellular interaction networks. The perturbation of any one molecule will inevitably perturb a subset of the global network. Therefore, to model the effect of drugs on cells it is important to be able to reconstruct networks of molecular interactions.

Knowledge of these networks should permit scientists to both discover new targets for therapeutic drugs as well as better understand the inherent toxic liabilities of these targets. Novel targets may be selected on the basis of their connectivity to pathways that have already been associated with diseases. For instance, if we know that several proteins are directly involved in the activation of T cells and we discover that a new protein is interacting with some of these proteins it is likely that the modulation of the new protein will have a direct impact on T cell activation. Many targets are currently being selected based on their network of interactions, and it is likely that in the future, as our understanding of these networks improves, this target selection strategy will become

commonplace. However, due to the long time involved in the development of drugs to targets, we have not yet seen drugs fully developed to targets selected using this approach.

Modeling the toxic consequences of drugs is an extremely challenging goal. Toxicity may arise due to a myriad of factors that include non specificity of the drug to inherently toxic protein targets. Interaction networks may prove valuable in our attempts to model the latter phenomenon. If we understand the local network of interactions a protein is involved in, it is likely that we will be able to better judge whether the modulation of the activity of the protein is likely to interfere with other critical pathways that would lead to toxic side effects.

One final example suggests how protein interactions are being considered for target selection. It is known that cancer cells typically contain multiple mutations with respect to the wild type cells from which they originate. These mutations allow the cancer cells to grow abnormally which ultimately leads to disease. Since each of these mutations may lead to the alteration of the activity of the associated protein, it is thought that if we understood the interactions of this protein we could select among these a therapeutic target whose modulation leads to cell death only in cells where the mutation is present. These targets could lead to non-toxic chemotherapies that are cytotoxic in mutated cells while leaving the wild type cells unaffected [1].

1.2 Methods for Measuring Protein Interactions

Several experimental techniques have been developed to directly probe protein interactions within a cell. The two-hybrid technique is based on the construction a bait and a prey protein that are fused to two halves of a transcription factor. If the bait and the prey protein interact the transcription factor is reconstituted and its activity is measured though the activation of the transcription of a reporter gene. This approach is a specific example of a general class of protein fragment complementation assays. Protein interactions may also be directly monitored using various co-purification techniques. A protein may be directly purified using a specific antibody or the protein may be tagged with another protein or a small molecule tag and then affinity purified. In all cases, if the selected protein interacts with other proteins these will be co-purified. The identity of the interacting partners may be deduced using mass spectrometry among other techniques. Protein microarrays are also emerging as a promising tool to directly observe protein interactions in a parallel fashion.

Several techniques have also been developed to reconstruct indirect protein couplings within the cell. These techniques do not necessarily predict which proteins physically interact but which ones are likely to be coupled within the cell. Couplings imply that if we modulate the activity of one gene it is likely to affect the activity of the coupled gene via interactions with other molecules. It is also assumed that perturbations of coupled genes are likely to lead to the same cellular phenotype.

Protein couplings may be deduced from the evolutionary history of pairs of proteins. If we find that two genes evolve in a correlated fashion it is likely that they are coupled within a cellular network. This correlation may be deduced by reconstructing the phylogenetic tree of two families of interacting proteins such as a family of ligands and receptors. Correlated evolution may also be deduced by finding pairs of genes that are always inherited together from the ancestral species. In the case of bacteria, they may

also be inherited as part of a block of genes, and therefore one may find the pair of genes coded near each other in multiple organisms. In a few cases, two genes in an ancestral species may be mutated to form a single gene in the daughter species.

Another form of molecular coupling may be inferred from the measurement of gene expression levels. Genes that are transcribed in a correlated fashion are inherently coupled within the cell. The activity of one is likely to be correlated with the activity of its partner since their mRNA concentrations are coupled. It is now possible to deduce these couplings in an efficient manner from the analysis of DNA microarray data.

A final technique for discovering proteins that are coupled within a cell is through the search for synthetic lethal mutations. In the case of yeast, for instance, it is known that most gene knock-outs lead to viable yeast strains. However, it is possible in many cases to find secondary knock-outs that are lethal in these strains. The pair of knock-outs that are lethal in combination but viable alone are termed synthetically lethal. Gene pairs that are synthetically lethal are clearly coupled within the cell even though are not necessarily physically interacting.

A synthesis of interaction data from all these methods may be performed to generate cell-wide interaction networks. These networks represent our current best guesses of the network properties of cellular couplings. The complexity of these networks is akin to that of many other networks found in nature and human societies. It is inevitable that our understanding of biology and medicine will undergo profound changes as our understanding of these networks improves.

2 Methods for Detecting Direct Protein Interactions

2.1 Two-Hybrid Method

GAL4 is a transcription factor that in the presence of galactose activates transcription of the GAL structural genes, which encode galactose metabolic proteins. The protein contains two domains, an activation domain and a DNA binding domain. In 1989 Fields and Song suggested that GAL4 hybrids could be used to report protein interactions [2]. The strategy consists of forming a fusion between protein A with the Gal4 activation domain and protein B with the GAL4 DNA binding domain. If protein A and B interact, the GAL4 activation domain and the GAL4 DNA binding domain are brought in proximity of each other and the reconstituted protein activates the transcription of a reporter gene which has been engineered to contain the GAL4 promoter. Thus, the presence of the reporter gene in yeast implies that protein A binds protein B.

Over the past few years this strategy has been scaled up so that it is now possible to efficiently measure thousands of binding events [3-6]. Typically the observation of a single binding event is not a reliable indication that two proteins are actually interacting [14]. This is due to the fact that both proteins are over-expressed in this assay and therefore the observed interaction may not be present in the wild type yeast where the concentrations may be significantly lower. Therefore a common strategy is to report only interactions that are observed more than once in duplicate screens.

Some additional limitations of the two-hybrid approach include the difficulty of elaborating interactions involving membrane proteins. To study membrane proteins one

must construct GAL4 fusions with only the extracellular or cytoplasmic domains of membrane proteins, adding an additional level of complexity to the assay.

The second limitation comes from the fact that to date most of the comprehensive two-hybrid screens have been conducted for yeast or bacterial proteins, but not yet for human proteins. For drug discovery applications, a map of human interactions is more valuable than that in yeast or bacteria, since these serve a limited role as model organisms. The technical difficulties of extending the two-hybrid approach to human proteins have been partially solved, and a comprehensive map of two-hybrid interactions between human proteins should become available in the near future.

2.2 Protein Fragment Complementation Assays

The two-hybrid approach represents a specific instance of a general approach developed to study direct protein-protein interactions termed protein fragment complementation assays (PCA) [8-10]. As in the two-hybrid approach, in PCAs half of a reporter protein is fused to protein A and the other half to protein B. If protein A and B interact, the two halves of the reporter protein reconstitute to restore its activity. The assay then reads out the activity of the reporter protein.

In one particular implementation of PCA, the reporter protein is dihydrofolate reductase (DHFR) [10]. DHFR is an essential enzyme that converts dihydrofolate into tetrahydrofolate, a methyl group shuttle required for the de novo synthesis of purines, thymidylic acid, and certain amino acids. Cells that do not have DHFR activity cannot survive in media depleted of nucleotides.

The strategy underlying this PCA assay consists of fusing half of DHFR to protein A and the other half to protein B, in cell lines where the activity of DHFR is knocked out. If protein A and B interact, the activity of DHFR is restored and the cells survive in nucleotide depleted media, conversely if protein A and B do not interact the cells die. It is critical to the success of this assay that the two halves of DHFR fused to the proteins A and B cannot fold unless the two query proteins interact.

This assay has been applied in Chinese hamster ovary cells lacking DHFR and multiple protein interactions have been probed. However it is possible to extend this assay one step further to probe the extent of interactions between two proteins [8]. To accomplish this one introduces one last step which consists of adding fluorescent methotrexate to the cells. Methotrexate is a known inhibitor of DHFR that binds with high affinity. The degree of interaction between two proteins with DHFR fractions fused may be monitored by fluorescence imaging of the methotrexate. This allows one to monitor the extent to which the interaction is perturbed as the cell is perturbed and hence map out the crosstalk between molecular pathways.

2.3 Co-purification techniques

One standard approach that may be utilized to map protein interactions is to tag a protein in the cell and then pull down the tagged protein together with other proteins bound to it. The identity of the interacting partners may then be revealed using various techniques among which are mass spectrometry. Two versions of this approach have recently been

implemented by groups that set out to map a large set of protein interactions in *Saccharomyces cerevisiae* [11,12].

In the first version the tandem affinity purification (TAP) tag was used [32]. This tag consists of two components: the first is used as a first purification tag which is then cleaved and the second tag is used for a subsequent purification step. The tag is built into the 3' end of the chosen gene using homologous recombination. This ensures that the tagged protein is expressed at native levels within the cell. Once the tagged protein has been affinity purified, the product is separated using one-dimensional SDS-PAGE. The identity of the proteins in the various bands are then determined using MALDI-TOF mass spectrometry.

In the second implementation genes are tagged with the Flag epitope [12]. In contrast to the previous approach, the genes are transiently overexpressed from the heterologous GAL1 or tet promeoters. The next steps are similar to the ones described above: the proteins are affinity purified, passed through an SDS-poluachrilamide gel, and then idenitifed using MS/MS fragmentation.

The accuracy of a procedure is difficult to determine, but may be approximated by checking the measured interactions against known ones. The coverage and accuracy of each approach may then be approximated. Although both approaches yielded comparable numbers of protein pairs, various studies suggest that purification of the TAP tag and native expression levels lead to fewer false positives using the first approach with respect to the second [13].

2.4 Protein Chips

Yet another approach to study protein interactions that has recently been developed consists of constructing protein chips. Protein chips are the protein counterpart of DNA chips which are widely used to measure gene expression levels. In a protein chip each spot consists of a different purified protein. By studying the binding of fluorescent molecules to these chips it is possible to reconstruct protein interaction patterns.

Protein chips are significantly more difficult to manufacture than DNA chips because of the intrinsic properties of proteins that allow them to bind nonspecifically to many surfaces. However, many of the technical difficulties have recently been surmounted and a nearly complete yeast protein chip has been assembled [15]. To attach proteins to the chip slide, each gene was fused to glutathione S-transferase polyhistidine (GST-HisX6). Each of 5800 different genes was then spotted onto a microarray using a standard microarrayer.

The advantage of protein chips over the previously mentioned methods for studying protein interactions is that they permit the observation of the binding of any molecule to a spot and not just another protein. It is therefore possible, for instance, to measure the binding of phosphoinositide (PI). The authors identified over 150 PI binding proteins that could be grouped into strong and weak binders according to the intensity of the fluorescent signal.

Difficulties associated with the manufacture of protein chips include purification of the proteins without their binding partners, misfolding of proteins on the glass slide and general difficulties in expressing large quantities of all yeast ORFS. Although protein chip technologies show great promise, it may yet take several years for their use to

become widespread as the technical difficulties surrounding their manufacturing and use are overcome.

3 Computational Methods for Detecting Interacting Partners

3.1 Evolutionary Evidence for Protein Couplings

During the course of evolution protein sequences may undergo multiple transformations such as point mutations, insertions, deletions and duplications. The ability to construct multiple alignments of proteins families allows one to partially reconstruct these transformations. The study of these evolutionary transformations of single protein families has become commonplace, and may be visualized with multiple sequence alignments or phylogenetic trees (e.g. [33]).

More recently however, as sequence databases have grown, it has become possible to study the correlation between the evolutionary relationships between two distinct protein families. One might imagine that if two proteins interact, the evolution of one might be correlated with the other. For instance, mutations that occur on a ligand might be compensated by mutations to its receptor in order to maintain the ligand-receptor binding affinity. This phenomenon has in fact been demonstrated in the case of chemokines and their associated receptors [16]. Therefore, by constructing phylogenetic trees of ligands and ligand receptors it is possible to partially reconstruct which ligand is likely to bid which receptor.

During the course of evolution proteins are not only mutated but occasionally deleted or horizontally transferred from one organism to another. Deletion or horizontal transfer events are particularly common in bacteria since the genomes are under selective pressure to remain compact. These phenomena may also be exploited to study protein couplings by reconstructing the correlated presence or absence of proteins within genomes. It has been observed that pairs of proteins that are lost or gained together during evolution tend to be functionally coupled [21].

One technique that has been used to study the correlated presence or absence of protein pairs across organisms involves the construction of phylogenetic profiles [17,18]. Phylogenetic profiles are simply a binary representation of the presence or absence of genes across species (see figure 1). The profile consists of a binary vector of N dimensions, where N is the number of fully sequenced genomes, whose entries are 1 if a protein family representative is present and zero otherwise. By identifying pairs of protein families that share similar phylogenetic profiles it is possible to identify which proteins are likely functionally coupled.

Finally, during the course of evolution genomes may undergo shuffling events that do not perturb individual gene sequences but perturb their order on the genome. In the case of eukaryotes where each gene possesses its own promoters these events are not likely to be very disruptive. However in bacteria where multiple genes belong to operons that are regulated by a single promoter, shuffling a genome may significantly affect the expression of genes. Therefore bacteria are under selective pressure to retain only

shuffling events that in large part maintain their operon structure. This assumption allows one to search for pairs of genes that have retained their chromosomal proximity in many species [19,20]. As in the case of proteins with matching phylogenetic profiles, these gene pairs are also likely to contain functionally associated genes. To date, however, this technique has only been applied to reconstructing couplings between bacterial genes. Recently this technique has been combined with that of phylogenetic profiles to study the evolutionary conservation of neighboring gene pairs [34]

3.2 Gene Fusions

As we discussed in the previous section, genes undergo multiple transformations during the course of evolution. One additional type of transformation involves the fusing of two genes into one. Although one may imagine that in most cases these mutations would be selected against as they might yield large and unstable proteins, in some cases they may be preserved because they enhance the function of the fusion protein by bringing into proximity two proteins with associated functions. One such example occurs in metabolism where a first protein generates a metabolic product that is acted upon by a second protein. If these two catalytic events are performed by two separate molecules the overall kinetics may be slower than if they are performed by a single molecule that combines them.

Using conventional sequence alignment techniques it is possible to systematically identify all fusion events that may have occurred between the genes of one organism [22,23,24]. This entails identifying a third protein that aligns, in a statistically significant fashion, to both the starting genes. However, this approach may also yield a significant number of spurious fusion events due to the fact that protein sequences are inherently modular and many modules are used hundreds of time throughout a genome. For example, finding two kinases that both align with a protein with two kinase domains is unlikely to represent a true fusion event since hundreds of kinases within the human genome contain multiple kinase domains.

3.3 Co-transcription

Human cells contain hundreds of transcription factors that bind to specific promoters within genes. Each gene typically contains multiple promoters and its transcription rates are therefore influenced by multiple factors. This scenario leads to a complex transcription network in which the mRNA levels of one gene are coupled to those of many others with either positive or negative correlations [35].

Since the development of DNA microarrays it has become possible to simultaneously measure most of the mRNA concentrations within cellular populations. As the cell is subject to perturbations the expression levels of many genes are altered. By monitoring the concentrations of most mRNAs as a cellular population is subject to these perturbations it is in principle possible to reconstruct many of the couplings between genes [25]. Coupled genes are likely to contain common promoters that are activated by some of the perturbations.

The Pearson correlation coefficient is usually computed to estimate the coupling between genes:

$$r(X,Y) := \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X_{ij}-\overline{X})(Y_{ij}-\overline{Y})}{\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X_{ij}-\overline{X})^2}\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(Y_{ij}-\overline{Y})^2}}$$

It is straightforward to compute all N^2 correlations between the N genes spotted on a DNA microarray. These are typically distributed according to a normal distribution bound between 1 and -1. The extreme tails of this distribution contain the highly positively and negatively correlated genes within the particular dataset that has been examined (see figure 2). These pairs are most likely involved in a common pathway or protein complex. For example, one typically finds that most of the protein components of the ribosome fluctuate in correlated fashion across experiments. A popular technique used to visualize the relationships between co-expressed genes involves the construction of hierarchical clusters [26].

DNA microarrays have become a powerful tool for pathway reconstructions because they permit the direct measurement of transcriptional couplings between genes. However, many interacting proteins are not co-transcribed even though it has been shown that interacting proteins tend to be more co-expressed than non-interacting proteins [14]. It is important therefore to use multiple techniques to gain a full understanding of protein couplings within cells.

3.4 Synthetic Lethal Mutations

One final technique that we review for establishing the interacting partners of proteins is the search fore synthetic lethal mutations. Following a systematic deletion of all yeast genes it was discovered that about 80% of genes are viable. In other words, yeast cells are able to grow and function when these genes are knocked-out one at a time [27]. However, in some cases when two individually viable genes are knocked out the cell dies. Genes that are viable individually, but essential when both are mutated are termed synthetically lethal.

The relationship between a pair of synthetically lethal genes is different than that identified by techniques that measure protein-protein physical interactions. Two genes that are synthetically lethal need not bind to each other, but must possess activities that are complementary so that the deletion of one may be buffered by the other. It has been observed that synthetically lethal pairs of proteins tend to participate within the same pathway [28]. Therefore these pairs represent interacting partners whose activities are coupled.

Recently, advances in robotics have permitted the systematic search for synthetic lethal pairs in *Saccharomyces cerevisiae* [28]. By crossing strains of yeast with single mutations it is possible to generate all possible double mutants. By screening these double mutants for viability, it is possible to identify all synthetic lethal pairs.

Initial results of such a screen have shown that on average each gene has about 20 synthetic lethal partners. This suggests that a complete network of synthetic lethal pairs would be highly interconnected, once again demonstrating that our current notions of

modular pathways may need to be rethought once interaction data is collected on a genome-wide scale.

4 Synthesizing Couplings Deduced form Multiple Methods into a Single Network

The data reported from experiments that probe direct protein-protein interactions has been catalogued within the Database of Interacting Proteins (DIP) [7]. Currently there are about 15,000 interactions between *Saccharomyces cerevisiae* proteins reported there, the organism within which comprehensive screens have been conducted most extensively (see Table I). When viewed as a network these relationships represent a comprehensive view of protein interactions within yeast, encompassing about two thirds of the yeast proteome. Although it is not known how many direct physical interactions between yeast proteins will ultimately be measured, it is assumed that each protein engages in only three or four interactions and that the current catalogue will not grow significantly in the future.

Databases have also been constructed to combine the evolutionary based methods for deducing protein couplings [29,30]. To date these have been mostly applied to bacteria where the existence of operons renders the methods more successful.

Finally, in the case of *Saccharomyces cerevisiae*, it has been shown that it is possible to combine protein couplings inferred from evolutionary methods and co-expression data to arrive at a comprehensive network of couplings between yeast genes [31]. Similar approaches are currently being applied to the reconstruction of human networks and being applied directly to the discovery of new drugs.

5 Summary

Protein targets consist of genes whose modulation has a therapeutic impact on disease. The difficulty of identifying these proteins is due to the difficulty of testing the modulation of proteins in humans where the drugs ultimately act. Typically these tests involve complicated clinical trials that involve the monitoring of hundreds or thousands of people over months or years at a cost of hundreds of millions of dollars. To enhance the success of these trials it is critical that one has a good disease model to select the therapeutic modality. These models typically consist of animals that manifest the disease or cell lines that reproduce certain aspects of the disease.

In the future, as robotics advances, it may be possible to subject these disease models to thousands of perturbations and systematically select out the ones that impact the disease state. For instance, it may be possible to subject a particular cell model to all possible gene deletions and monitor which ones impact the disease phenotype. However, currently it is not usually feasible to conduct such genomic screens and it is therefore imperative to design sub-genomic perturbation experiments. It is in the design of these experiments that protein interaction maps are proving valuable. Using protein interaction networks it is possible to expand a small sets of genes that are likely to affect the modeled disease states into a much larger set that are linked to these. Experiments

that set out to demonstrate that these expanded lists of genes are involved in the disease state should be highly enriched for positive results compared to genome-wide assays.

In the future, as static interaction networks are replaced by more sophisticated quantitative models it may become possible to accurately predict the outcome of many experiments *in sislico*. The realization of this goal should have a dramatic impact on the way drug targets are discovered. The next few decades will undoubtedly see a great deal of research devoted to this long term goal.

Bibliography

1. HARTWELL LH, SZANKASI P, ROBERTS CJ, MURRAY AW, FRIEND SH. Integrating genetic approaches into the discovery of anticancer drugs. *Science* (1997) 278:1064-1068.

2. FIELDS S, SONG O. A novel genetic system to detect protein-protein interactions. *Nature* (1989) 340:245-246.

3. UETZ P, GIOT L, CAGNEY G *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* (2000) 403:623-7.

4. RAIN JC, SELIG L, DE REUSE H *et. al.* The protein-protein interaction map of *Helicobacter pylori*. *Nature* (2001) 409:211-5.

5 ITO T, CHIBA T, OZAWA R *et. al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* (2001) 98:4569-74.

6 ITO T, TASHIRO K, MUTA S *et. al.* Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* (2000) 97:1143-7.

7 XENARIOS I, SALWINSKI L, DUAN XJ *et. al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* (2002) 30:303-5.

8 REMY I, MICHNICK SW. Visualization of biochemical networks in living cells. *Proc Natl Acad Sci U S A* (2001) 98:7678-83.

9 REMY I, MICHNICK SW. Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. *Proc Natl Acad Sci U S A* (1999) 96:5394-9.

10 PELLETIER JN, CAMPBELL-VALOIS FX, MICHNICK SW. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc Natl Acad Sci U S A* (1998) 95:12141-6.

11 GAVIN AC, BOSCHE M, KRAUSE R *et. al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* (2002) 415:141-7.

12 HO Y, GRUHLER A, HEILBUT A *et. al*. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* (2002) 415:180-3.

13 VON MERING C, KRAUSE R, SNEL B *et. al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* (2002) 417:399-403.

14 DEANE CM, SALWINSKI L, XENARIOS I, EISENBERG D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* (2002) 1:349-56.

15 ZHU H, BILGIN M, BANGHAM R *et. al.* Global analysis of protein activities using proteome chips. *Science* (2001) 293:2101-5.

16 GOH CS, BOGAN AA, JOACHIMIAK M, WALTHER D, COHEN FE. Co-evolution of proteins with their interaction partners. *J Mol Biol*. (2000) 299:283-93.

17 PELLEGRINI M, MARCOTTE EM, THOMPSON MJ, EISENBERG D, YEATES TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* (1999) 96:4285-8.

18 VERT JP. A tree kernel to analyse phylogenetic profiles. *Bioinformatics* (2002) 18 Suppl 1:S276-84.

19 OVERBEEK R, FONSTEIN M, D'SOUZA M, PUSCH GD, MALTSEV N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* (1999) 96:2896-901.

20 DANDEKAR T, SNEL B, HUYNEN M, BORK P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* (1998) 23:324-8.

21 HUYNEN MA, BORK P. Measuring genome evolution. *Proc Natl Acad Sci USA* (1998) May 26;95(11):5849-56.

22 MARCOTTE EM, PELLEGRINI M, NG HL *et. al*. Detecting protein function and protein-protein interactions from genome sequences. *Science* (1999) 285:751-3.

23 ENRIGHT AJ, ILIOPOULOS I, KYRPIDES NC, OUZOUNIS CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* (1999) 402:86-90.

24 YANAI I, DERTI A, DELISI C. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci USA* (2001) 98:7940-5.

25 HUGHES TR, MARTON MJ, JONES AR *et. al.* Functional discovery via a compendium of expression profiles. *Cell* (2000) 102:109-26.

26 EISEN MB, SPELLMAN PT, BROWN PO, BOTSTEIN D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* (1998) 95:14863-8.

27 WINZELER EA, SHOEMAKER DD, ASTROMOFF A *et. al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* (1999) 285:901-6.

28 TONG AH, EVANGELISTA M, PARSONS AB *et. al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* (2001) 294:2364-8.

29 PELLEGRINI M, THOMPSON M, FIERRO J, BOWERS P. Computational method to assign microbial genes to pathways. *J Cell Biochem* (2001) Suppl 37:106-9.

30 MELLOR JC, YANAI I, CLODFELTER KH, MINTSERIS J, DELISI C. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res* (2002) 30:306-9.

31 MARCOTTE EM, PELLEGRINI M, THOMPSON MJ, YEATES TO, EISENBERG D. A combined algorithm for genome-wide prediction of protein function. *Nature* (1999) 402:83-6.

32 RIGAUT G, SHEVCHENKO A, RUTZ B *et. al*. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* (1999) 17:1030-2.

33 HIGGINS D., THOMPSON J., GIBSON T. *et. al*. CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*(1994) 22:4673-4680.

34 ZHENG Y, SZUSTAKOWSKI JD, FORTNOW L, ROBERTS RJ, KASIF S. Computational identification of operons in microbial genomes. *Genome Res*. (2002) 12:1221-30.

35 LEE TI, RINALDI NJ, ROBERT F *et. al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* (2002) 298:799-804.

Table I – Statistics from the Database of Interacting Proteins a compendium of experimental measurements of direct protein-protein interactions

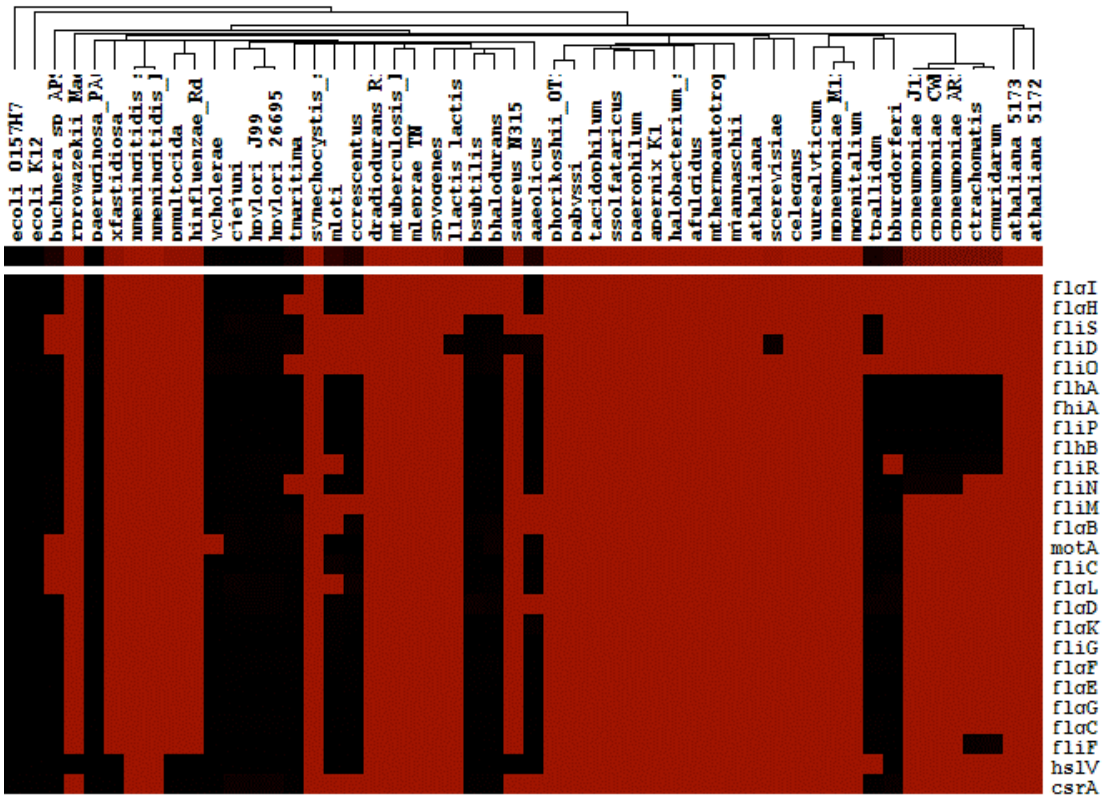
ORGANISM	PROTEINS	INTERACTIONS	#Exp	#Int
Saccharomyces cerevisiae (budding yeast)	4711	14941	1	13145
			2	1151
			3	350
			4	146
			5+	149
Helicobacter pylori	710	1415		
Homo sapiens (Human)	687	717	1	557
			2	103
			3	29
			4	17
			5+	11
Escherichia coli	269	286	1	192
			2	49
			3	24
			4	11
			5+	10
Mus musculus (house mouse)	177	97	1	81
			2	13
			3	3

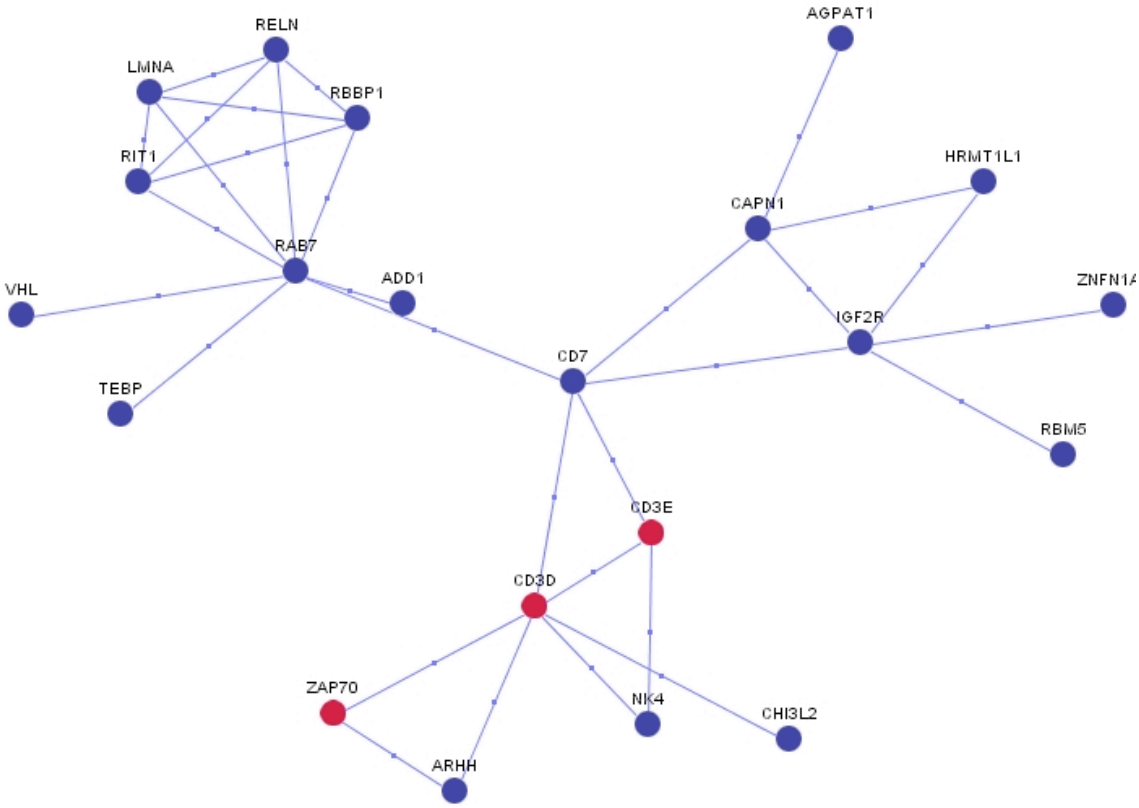
Figure 1

A hieraerchical cluster of *Eschericia coli* phylogenetic profiles. The cluster contains mostly genes known to be part of the flagella machinery. Red indicates that a homolog of the genes is present in the organism shown on the top, and black that it is absent. The cluster was constructed using the program Cluster and TreeView by Michael Eisen.

Figure 2

Co-transcribed genes deduced from a dataset of leukocyte cancers. We represent genes that are co-expressed above a cutoff value as linked. The red genes are known components of the T cell receptor: the epsilon and delta subunits of the CD3 T-cell receptor complex and ZAP70, a zeta chain associated kinase. The blue genes are additional genes co-expressed with these three. These linkage representations are useful for selecting novel genes that may modulate the activity of the T-cell receptor.


ecoli 0157H7
ecoli K12
buchnera sp AP
rprowazekii Ma
paeruginosa_PA
xfastidiosa
mmeningitidis
mmeningitidis_
pmultocida
hinfluenzae_Rd
vcholerae
cjejuni
hpylori J99
hpylori 26695
tmaritima
synechocystis_
mloti
ccrescentus
dradiodurans R
mtuberculosis_
mleprae TM
spyogenes
llactis lactis
bsubtilis
bhalodurans
saureus N315
aaeolicus
phorikoshi_OT
pabyssi
tacidophilum
ssolfataricus
paerophilum
apernix K1
halobacterium_
afulgidus
mthermoautotro
mjannaschii
athaliana
scerevisiae
celegans
uurealyticum
mpneumoniae_M
mpenitalium
tpallidum
bburgdorferi
cpneumoniae J
cpneumoniae CW
cpneumoniae AR
ctrachomatis
cmuridarum
athaliana 5173
athaliana 5172
flgI
flgH
fliS
fliD
fliO
flhA
fhiA
fliP
flhB
fliR
fliN
fliM
flgB
motA
fliC
flgL
flgD
flgK
fliG
flgF
flgE
flgG
flgC
fliF
hslV
csrA


RELN
LMNA
RBBP1
RIT1
RAB7
VHL
ADD1
TEBP
AGPAT1
HRMT1L1
CAPN1
ZNFN1A
IGF2R
CD7
RBM5
CD3E
CD3D
ZAP70
NK4
CHI3L2
ARHH