Progress in Drug Research, Vol. 64
(Helena I. Boshoff and Clifton E. Barry III, Eds.)

The protein network as a tool for finding novel drug targets

By Michael Strong
and David Eisenberg

Howard Hughes Medical Institute,
UCLA-DOE Institute of Genomics and
Proteomics,
University of California Los Angeles,
Los Angeles, California, USA
<mnstrong@gmail.com>

Abstract

Proteins are often referred to as the molecular workhorses of the cell since they are responsible for the majority of functions within a living cell. From the generation of energy, to the replication of DNA, proteins play a central role in most cellular functions. Because of their importance to cellular viability, proteins are commonly the target of therapeutic drugs, ranging from antimicrobial to anticancer drugs. With the rise of drug resistant and multi-drug resistant forms of many diseases, it has become increasingly important to develop new strategies to identify alternative drug targets. One such strategy arises from the analysis of protein networks. Protein networks help define individual proteins within the context of all other cellular proteins. In this chapter we discuss methods for the identification and analysis of genome-wide protein networks, and discuss how protein networks can be used to aid the identification of novel drug targets.

1 Protein linkages

Proteins can function together in many ways, ranging from direct physical associations among proteins in a complex, to transient interactions that occur among members of certain protein pathways. Proteins can also function as non-interacting members of the same pathway. As a result, it has been of great interest to develop methods to identify these protein associations, or protein linkages, on a genome-wide basis [1]. The detection of protein linkages has been aided by advances in both biochemical [2–6] and computational methods [7–15], which have yielded valuable insight into the underlying architecture of cellular networks [16–20].

2 Biochemical methods to identify protein-protein interactions

2.1 Yeast two-hybrid assay

One of the most widely used methods for identifying physically interacting proteins is the yeast two-hybrid assay (Y2H) [21]. The yeast two-hybrid assay enables the detection of physically interacting proteins, by exploiting the modular organization of transcriptional activators. Transcriptional activators contain two domains, a DNA binding domain and a

transcriptional activator domain, which together can initiate transcription of a target gene. When separated, however, these domains cannot initiate transcription on their own, unless they are brought into close proximity by additional factors.

In the yeast two-hybrid assay, the DNA-binding domain (DBD) of a transcriptional activator is fused to one protein of interest. This fusion protein is known as the 'bait' protein. The transcriptional activating domain (AD) is fused to another protein, known as the 'prey' protein. If there is a physical interaction between the 'bait' protein and the 'prey' protein, then the DNA-binding domain and the transcriptional activating domain come into close proximity and activate a specific reporter gene [21]. If the bait protein and the prey protein do not interact, however, then the DNA-binding domain and the transcriptional activating domain do not come into close proximity, and thus do not activate the reporter gene. This method has been scaled up to enable the high-throughput detection of genome-wide protein–protein interactions [22], and has greatly aided the identification of protein interactions in organisms including yeast [2, 3], *C. elegans* [23], *Drosophila* [24], and humans [25].

2.2 Co-immunoprecipitation method

Another widely used method for detecting protein–protein interactions is the co-immunoprecipitation method (Co-IP) [26]. In the Co-IP method, an antibody is made to target a particular protein of interest. The antibody is then added to a mixture of proteins, often comprising the total cellular lysate of a particular cell type, and allowed to bind to the target protein. If the target protein interacts with additional proteins, then protein–protein interactions can be identified by capturing the antibody and all attached proteins on a solid support. After washing unbound proteins away, the antibody and attached proteins can be eluted and analyzed by a variety of methods ranging from gel electrophoresis to mass spectrometry. Proteins that interact with the target protein are identified in this manner [26].

2.3 Co-affinity purification coupled with mass spectrometry

Variations of the Co-IP method have also been employed to detect physically interacting proteins, including the co-affinity purification (Co-AP) method coupled with mass spectrometry (AP-MS) [4, 5]. In this strategy, a specific target protein is tagged with an affinity tag, expressed with other cellular proteins, and affinity purified. Protein–protein interactions are detected by the co-purification of additional proteins with the tagged protein. Mass spectrometry is then used to identify interacting proteins. This application has been applied to investigate the proteome of *Saccharomyces cerevisiae* [4, 5], where it has enabled the identification of hundreds of protein complexes [4, 5].

2.4 Protein–protein interaction databases

To date, over 50,000 protein–protein interactions have been reported in the literature and catalogued into various databases [27]. Among these databases are the Database of Interacting Proteins (DIP) [28], the Biomolecular Interaction Network Database (BIND) [29], and the Molecular Interactions Database (MINT) [30]. Additionally, a number of web servers have arisen to catalog both known and putative protein pathways. These servers include the Kyoto Encyclopedia of Genes and Genomes (KEGG) [31], the Encyclopedia of *E. coli* Genes and Metabolism (EcoCyc) [32], and the Munich Information Center for Protein Sequences (MIPS) [33]. Together these databases and web servers provide a useful source for investigating protein–protein interactions in organisms ranging from *E. coli* to human.

3 Computational methods to identify protein linkages

In addition to biochemical methods to identify linked proteins, a number of computational methods have been developed to identify functionally linked proteins, including the Rosetta Stone [8], Phylogenetic Profile [11], conserved Gene Neighbor [14, 15], and Operon/Gene Cluster [13, 34] methods. Each of these methods utilizes genomic sequence information garnered from genome sequencing efforts. Currently there are over 300

completed genomes available [35, 36], and over 1,000 ongoing genome sequencing efforts [35]. Together these efforts provide us with a tremendous amount of information regarding not only the genetic blueprint of hundreds of organisms, but also facilitate the computational inference of protein linkages and protein networks.

3.1 Rosetta Stone method

The Rosetta Stone method provides a means for inferring protein linkages based on genomic analyses [8]. The Rosetta Stone method identifies individual genes in one genome that occur as a single fusion gene in another genome. For example, the *leuC* and *leuD* genes of *Mycobacterium tuberculosis* (*Mtb*) occur as two separate genes [37], but in *Schizosaccharomyces pombe* these two genes occur as a single fused gene. Based on this observation, it can be inferred that the *M. tuberculosis leuC* and *leuD* genes are 'functionally linked'. Functionally linked genes may represent genes that encode members of a common protein complex, a common protein pathway, or proteins that serve related functions within the cell [1]. While the *leuC* and *leuD* example demonstrates a Rosetta Stone linkage between two genes of known function (both genes are involved in leucine biosynthesis), many Rosetta Stone linkages involve uncharacterized proteins [8].

3.2 Phylogenetic Profile method

A second method for inferring protein linkages is the Phylogenetic Profile method [11]. The Phylogenetic Profile method identifies genes that occur in a correlated manner across many genomes, specifically identifying genes that are present or absent in a correlated manner [11]. For example, the *fliC* and *fliG* genes of *E. coli* share similar Phylogenetic Profiles. Both *fliC* and *fliG* are present in genomes of flagellated motile bacteria, while both proteins are absent in genomes of non-motile bacteria. We might expect that genes that participate in a shared biochemical pathway or protein complex would share similar phylogenetic profiles.

3.3 Conserved Gene Neighbor method

A third method for inferring protein linkages is the conserved Gene Neighbor method [14, 15]. This method identifies genes that tend to be located in close chromosomal proximity in multiple genomes. For example, the *E. coli otsA* and *otsB* genes are both involved in trehalose biosynthesis, and are located in close chromosomal proximity in a number of genomes including *E. coli*, *S. typhi*, and *M. loti*. The close chromosomal positioning of genes across many genomes is a common feature of genes in bacterial operons, and suggests related functions. This is also observed in eukaryotic organisms, although to a lesser extent.

3.4 Operon/Gene Cluster method

The Operon method [13], also referred to as the Gene Cluster method [38], utilizes information from a single genome to identify putative operon members based on the distance between adjacent genes in the same orientation [13]. Genes that are separated by minimal intergenic distances are more likely to belong to common operons than genes separated by larger distances [10, 12, 39]. This method has been applied to identify linked genes in organisms ranging from *E. coli* [12] to *M. tuberculosis* [13], and this method is particularly useful in instances where no identifiable gene homologs are present. To date, most genome sequencing ventures have identified genes that are completely unique to a particular organism, and in these cases, the Operon/Gene Cluster method may be particularly useful for assigning putative function or linking uncharacterized genes to characterized genes.

4 Databases of inferred protein linkages

Collectively, the described computational methods provide a powerful tool to infer protein linkages, which can then be used to construct genome-wide protein networks. As the number of completed genomes continues to increase, these methods are likely to become more powerful. Currently the ProLinks Database [38] contains inferred protein linkages for over 160

sequenced genomes, and includes over 17 million high confidence linkages [38, 40] identified by the Rosetta Stone, Phylogenetic Profile, conserved Gene Neighbor, and Operon/Gene Cluster methods. Another useful database of inferred protein linkages is the EMBL STRING server [41].

5 Protein networks

Biochemical and computational methods have greatly facilitated the identification of protein linkages on a genome-wide scale. The next question we can ask is "How are these protein linkages organized on a genome-wide scale?" This question can be answered by the construction and analysis of protein networks. Protein networks provide a useful graphical method to investigate the connectivity of individual proteins, as well as sets of proteins [16–18]. Figure 1 depicts a protein network centered on the human cellular tumor antigen p53. p53 is an important tumor suppressor gene [42] that is frequently mutated or inactivated in human cancer cells [43]. As a result, this protein has been thoroughly studied at both the cellular and molecular level.

Figure 1a shows a list of proteins that p53 has been found to physically interact with, as retrieved from the Database of Interacting Proteins [28]. p53 interacts with a number of proteins, including other important cancer-related proteins such as the Breast cancer type 1 (BRCA1) and type 2 (BRCA2) susceptibility proteins, as shown in Figure 1a. Figure 1b depicts the same interactions listed in Figure 1a, but in this case the data are represented as a protein network. In the network, each protein is represented as a circular 'node', and each interaction is indicated as a connecting line, better known as an 'edge'. The p53 protein serves as the central node in this network. The network depicts proteins that interact directly with p53, as well as proteins that are linked by two edges. Protein networks facilitate the analysis of protein linkages and provide a useful graphical interface for analyzing and interpreting large amounts of data.

While the p53 protein network of Figure 1b was constructed using experimentally identified protein–protein interactions, protein networks can also be constructed using computationally inferred protein linkages [38]. Such methods have the advantage of providing information regarding or-

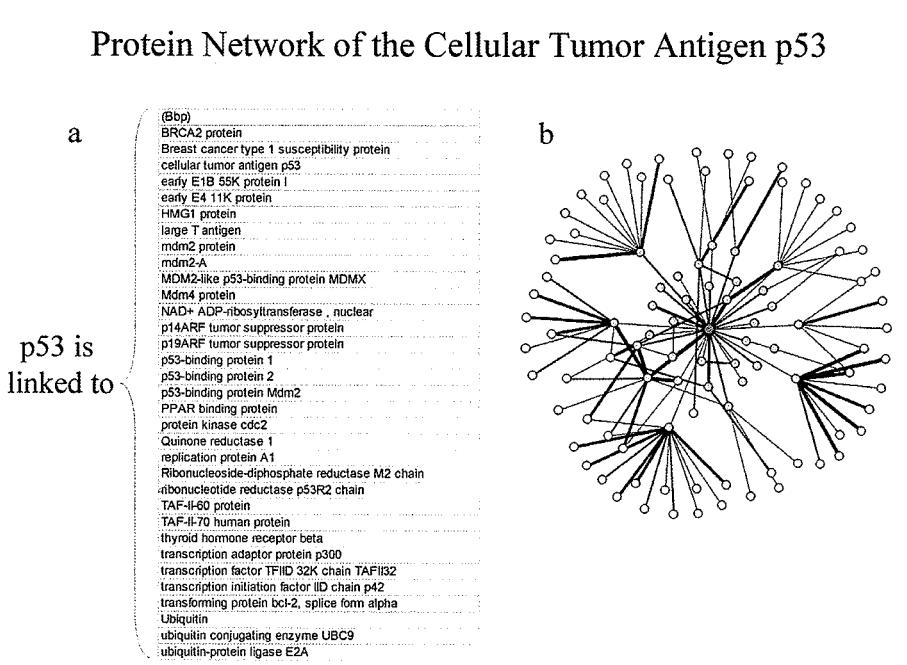


Figure 1.
Protein network of the cellular tumor antigen p53. a) List of proteins that p53 interacts with, as retrieved from the Database of Interacting Proteins [28]. b) p53 protein network. P53 serves as the central node in this network, with 1st and 2nd shell nodes depicted.

ganisms in which extensive biochemical or genetic experiments have not been done. Figure 2 depicts a computationally inferred protein network centered on the yeast prion protein Sup35. This network was constructed using a combination of the Phylogenetic Profile method (PP), the Rosetta Stone method (RS), the conserved Gene Neighbor (GN) method, and the Operon/Gene Cluster (GC) method [38].

The yeast prion protein, Sup35, has been shown to exhibit properties of prion-like infectivity [44, 45], resulting from the formation of amyloid-like fibrils [46–49]. The Sup35 network reveals a number of linkages to proteins involved in transcription and translation activities, which may be related to the natural cellular function of Sup35. The use of computationally inferred protein networks, such as the Sup35 network, as well as biochemical-based protein networks, such as the p53 network, may help us

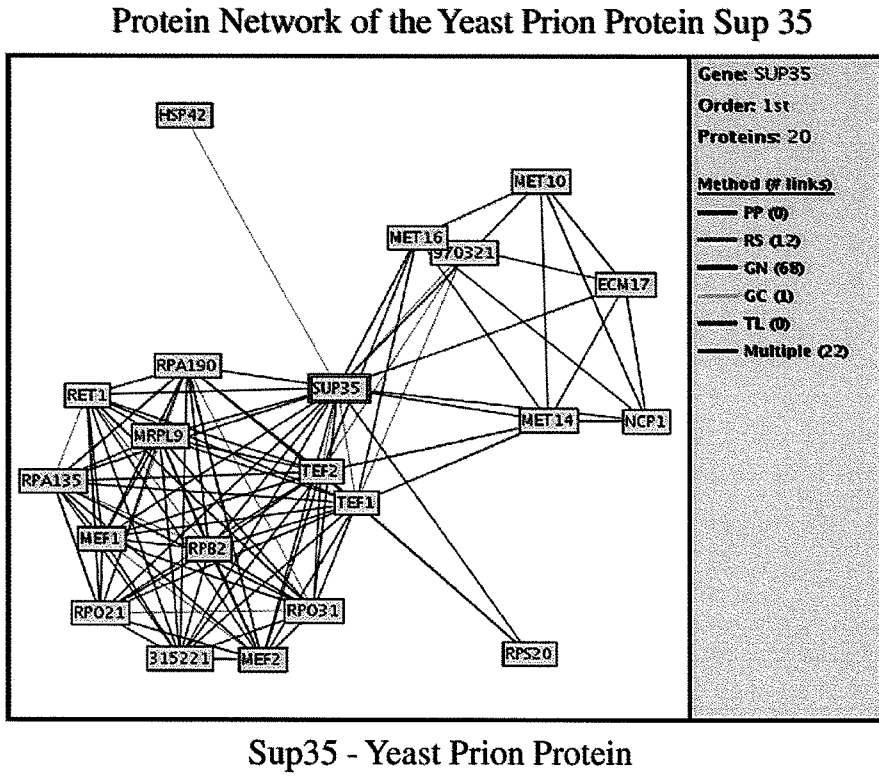


Figure 2.
Protein network of the yeast prion protein Sup35. Linkages indicated in this type of network are inferred by the Phylogenetic Profile (PP), Rosetta Stone (RS), conserved Gene Neighbor (GN), and Operon/Gene Cluster (GC) computational methods.

better understand the molecular framework in which normal and disease-associated proteins function, and in turn may suggest new strategies to combat a variety of diseases.

Figures 1 and 2 represent somewhat simplified protein networks with only the 1st and 2nd shell nodes depicted. Many protein networks, however, exhibit higher complexity, as shown in Figure 3. In some cases, protein networks comprise hundreds or even thousands of linkages. While the classical method of protein network representation has relied on the node and edge type network (Fig. 3), recent work has demonstrated useful advantages of matrix-represented protein networks [19, 20, 50].

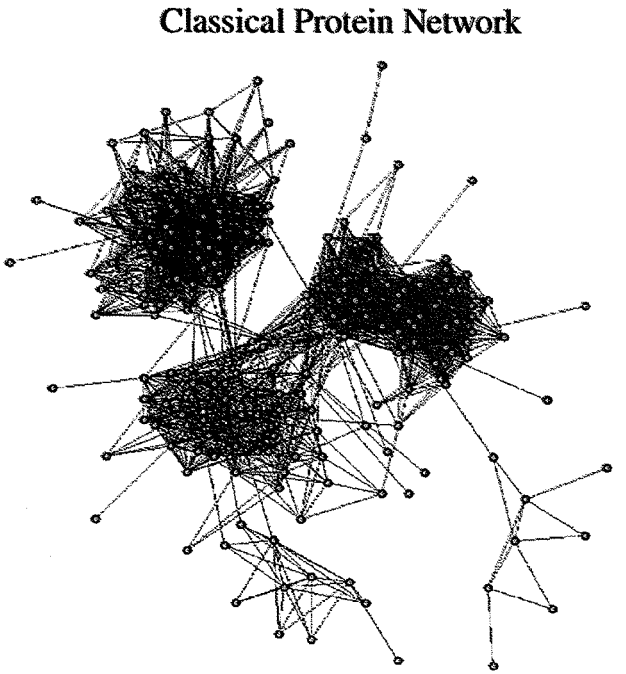


Figure 3.
Classical protein network depicting *M. tuberculosis* protein linkages. Figure adapted from Strong et al. [50].

3.1 Matrix-represented protein networks – genome maps

An alternative approach to represent genome-wide protein networks is shown in Figure 4. In this approach, each linked pair of proteins is indicated as a single point on a two dimensional matrix, corresponding to the position of the genes on the chromosome [50]. Each axis of the graph represents a monotonically ordered list of genes, starting at the origin of replication and proceeding along the chromosome. The *M. tuberculosis* genome has approximately 4,000 genes, as indicated on the x and y axis of the matrix in Figure 4c. Each point on this graph indicates a computationally inferred protein linkage between two proteins [50]. Figure 4a depicts a zoomed in region of the map, representing only the first 50 genes. The point at coordinate x=1, y=5 represents a linkage between the 1st gene on the *M. tuberculosis* chromosome (Rv0001, dnaA) and the 5th gene on

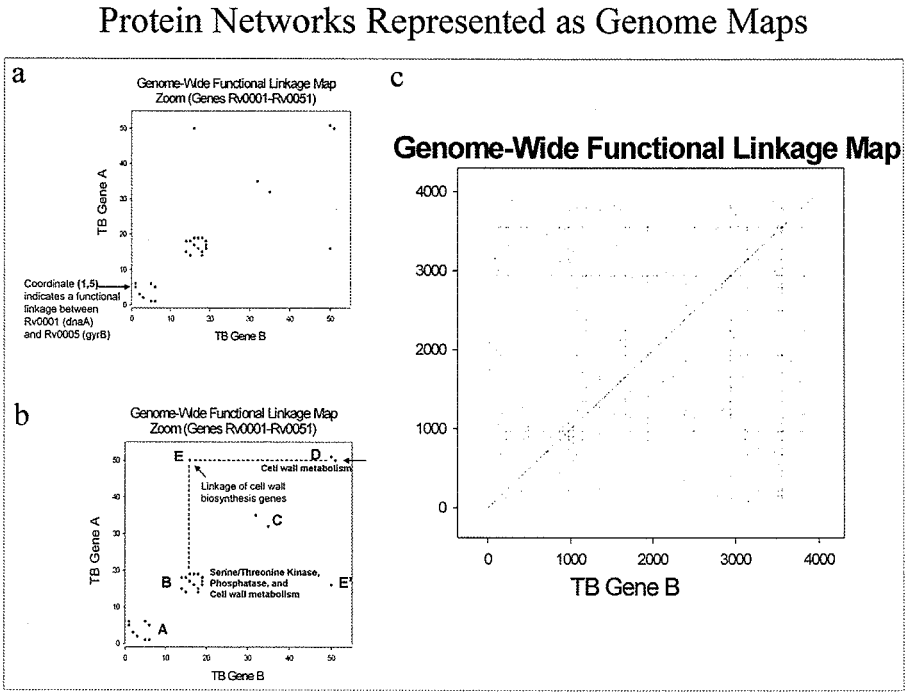


Figure 4.
Genome-wide functional linkage map. a) Zoomed in region of the genome-wide functional linkage map depicting the first 50 genes. Genes are organized according to the order on the chromosome. Each 'point' on the matrix represents a pair of functionally linked genes, for instance the point at coordinate x=1, y=5 indicates a linkage between the first gene, Rv0001 (dnaA) and the fifth gene, Rv0005 (gyrB). b) Functional categories of some of the proximal genes. c) Complete genome-wide functional linkage map depicting nearly 10,000 high confidence functional linkages in *M. tuberculosis*. Figure adapted from Strong et al. [50].

the chromosome (Rv0005, gyrB). Both these genes are involved in DNA replication or repair.

The representation of protein networks as two dimensional genome maps reveals certain characteristics that are not observable using traditional node and edge protein networks. Since information regarding chromosomal organization is maintained in the genome maps, we can analyze protein connectivity in relation to genome organization. One feature that is readily apparent in the genome map of Figure 4 is the local connectivity of genes that are located in close chromosomal proximity [50]. In many

cases these clusters of highly connected genes correspond to known or putative operons. Often these clusters contain genes that perform related cellular functions. For instance in Figure 4b, cluster A, most of the genes are involved in DNA replication or repair. In cluster B, there are two genes encoding serine threonine kinases, one phosphatase, and two cell wall metabolism genes. Due to the functional connectivity among the genes of this region, it can be hypothesized that the genes of this cluster participate in a cell wall signaling cascade [50]. This hypothesis was further supported by the presence of a putative peptidoglycan-sensing domain on one of the serine-threonine kinase proteins [51].

The Genome-wide Functional Linkage Map represented in Figure 4c contains approximately 10,000 high confidence protein linkages, inferred by two or more computational methods. To further facilitate the analysis of these protein networks Strong et al. also developed a method to hierarchically cluster the genes of the matrix, based on the similarity of the functional linkage profiles [50]. A functional linkage profile indicates all genes a particular gene is linked to, represented as a bit vector. A '1' in the bit vector indicates a protein linkage and a '0' indicates the absence of a linkage. In the hypothetical example shown in Figure 5a, Gene A is linked to Gene B, Gene C, and Gene D, as indicated by the '1's in the profile. Profiles are then clustered using a hierarchical clustering algorithm, bringing together genes that share similar functional linkage profiles.

The resulting clustered map, shown in Figure 5b, reveals important characteristics of protein network connectivity and hierarchy. Many of the genes cluster into distinct modules, participating in related cellular functions [50]. Some of these modules correspond to protein pathways or complexes, while others contain genes that serve related cellular functions. Some of the functional modules are indicated in Figure 5b. Figure 5c depicts a zoomed-in region of the clustered map, indicated by the black square. Functional modules in this region correspond to genes involved in detoxification, polyketide synthesis, energy metabolism, and the degradation of fatty acids. This example illustrates how hierarchical clustering of genomic maps can enable the rapid identification of functional modules on a genome-wide basis [50].

Figure 6 shows ten representative clusters of the hierarchically clustered map. In some cases, the gene clusters can be used to infer protein

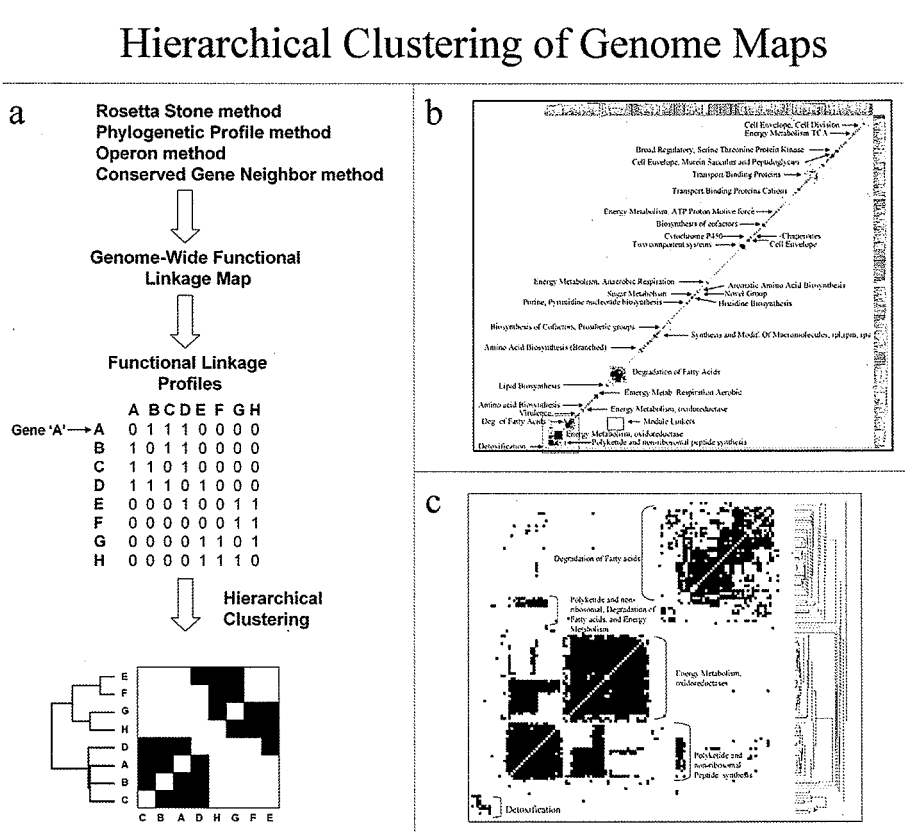


Figure 5.
Hierarchical clustering of the genome-wide functional linkage map. a) Outline of the method. b) Hierarchical clustering reveals the inherent modularity of the *M. tuberculosis* genome. c) Representative *M. tuberculosis* functional modules. Figure adapted from Strong et al. [50].

function for uncharacterized genes. In Figure 6a, a group of chaperone proteins cluster with a non-annotated gene, Rv2372c. Based on this observation, it can be inferred that Rv2372c has a function associated with that of the chaperones of this cluster. In Figure 6b, a number of genes involved in the synthesis and modification of polysaccharides cluster with the uncharacterized gene Rv0127. Based on this clustering, Rv0127 is hypothesized to be involved in polysaccharide synthesis or modification. In other cases, clusters contain a large percentage of non-annotated genes (Fig. 6d–j). These clusters may suggest previously uncharacterized mod-

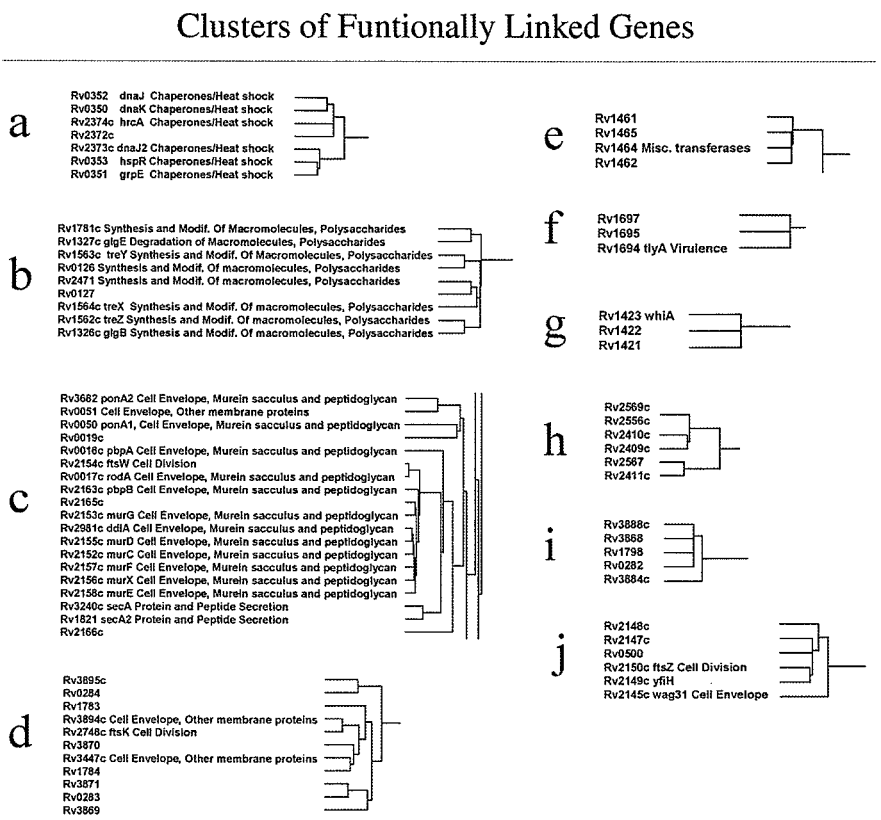


Figure 6.
Representative clusters of functionally linked genes. Gene clusters can aid the inference of gene function for uncharacterized genes as well as can identify novel groups of genes that may function together as a unit. Figure adapted from Strong et al. [50].

ules, possibly corresponding to members of common pathways or complexes, yet to be characterized. A more comprehensive understanding of the modularity of genome-wide protein networks in human pathogens may enable researchers to better devise strategies to combat the pathogenic effects that certain modules are responsible for.

Gene expression analyses have also become an essential tool to identify genes that play important roles during disease states or during infection. While gene expression analyses alone can be used to identify important genes, the examination of gene expression within the context of protein

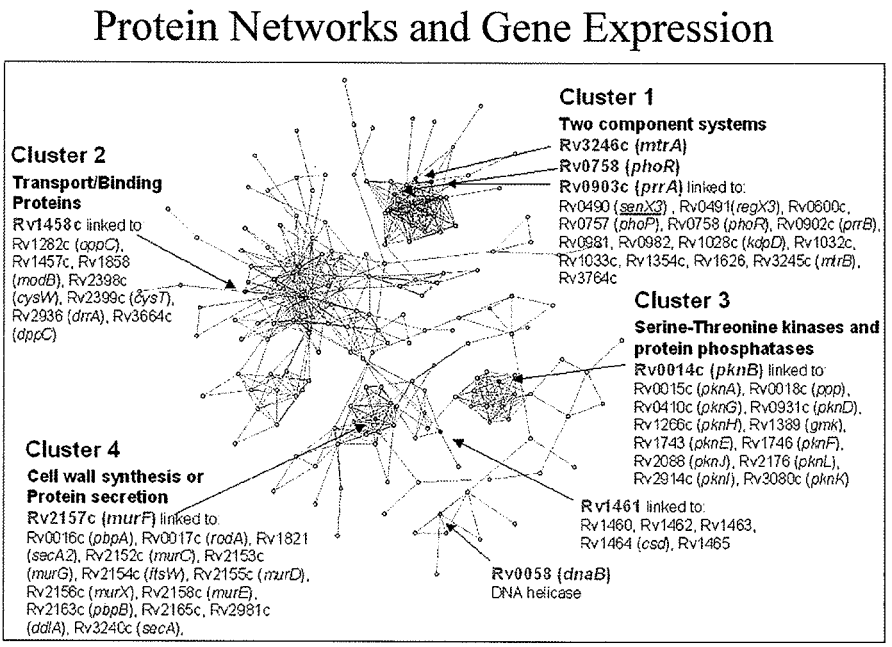


Figure 7.
Examination of gene expression patterns within the context of protein networks. Upregulated genes are indicated by the arrows. Figure adapted from Rachman et al. [52].

networks may further help us to understand the mechanisms by which certain systems are triggered during disease states or infection [52]. Figure 7 shows an example of *M. tuberculosis* gene expression profiling within the context of computationally inferred protein networks. In this case, *M. tuberculosis* genes that are upregulated during macrophage infection are indicated by arrows. Analyses such as these may aid the identification of modules that are important during infection, and may be useful in narrowing the field of potential drug targets.

6 Drug targets

One of the major challenges confronting many branches of infectious disease control is the emergence of drug resistant strains of many viral and

bacterial pathogens [53]. Amplifying this concern is the emergence, in some cases, of multi-drug resistant strains [54]. As a result, there is a dire need for the identification of effective, alternative drug targets that may be used to combat these pathogens as they become resistant to current drugs. Often, drug resistance emerges as a result of specific amino acid alterations in targeted proteins [55]. In some cases, these mutations render drugs ineffective, while in other cases they decrease the efficiency of the drug. Resistance to penicillin, for example, is associated with specific amino acid mutations in the penicillin binding proteins [56].

While many drugs target a specific protein, the resulting activity of a drug is often the disruption of a particular cellular function, pathway, or complex. For example, fluoroquinolones inhibit the DNA unwinding activity of the gyraseAB complex, penicillin inhibits cell wall biosynthesis by targeting the penicillin binding proteins, rifampin inhibits the transcriptional activity of the RNA polymerase complex by targeting the RpoB protein, and streptomycin inhibits protein synthesis which can be alleviated by mutations in the *rpsL* gene [57]. In effect, each drug, by targeting a specific protein or small group of proteins, inhibits or disrupts an important cellular pathway, complex, or function. As protein targets become resistant, it may be useful to target other members of the same pathway or complex, as well as proteins that serve related cellular functions. In these cases, protein networks can be useful for the identification of new drug targets that are linked directly or indirectly to current drug targets.

Figure 8 shows computationally inferred protein networks involving four anti-tuberculosis drug targets, RpoB (the target of Rifampin), KasA (a target of Isoniazid), GyrA (the target of Fluoroquinolone drugs), and RpsL (the target of Streptomycin). Each of these networks was generated using the ProLinks server [38]. In each of these cases, we see that proteins of similar cellular function are linked. In the case of RpoB, the Rifampin drug target, there are linkages to other transcription related proteins such as RpoC (the RNA polymerase beta' subunit) and NusG (the transcription antitermination protein), as well as a number of ribosomal proteins.

In the GyrA protein network, GyrA is linked to GyrB (the other member of the DNA gyrase AB complex), the DNA replication initiator DnaA, the DNA replication and repair protein RecF, and the DNA polymerase III protein DnaN. GyrA is also linked to the uncharacterized gene Rv0007.

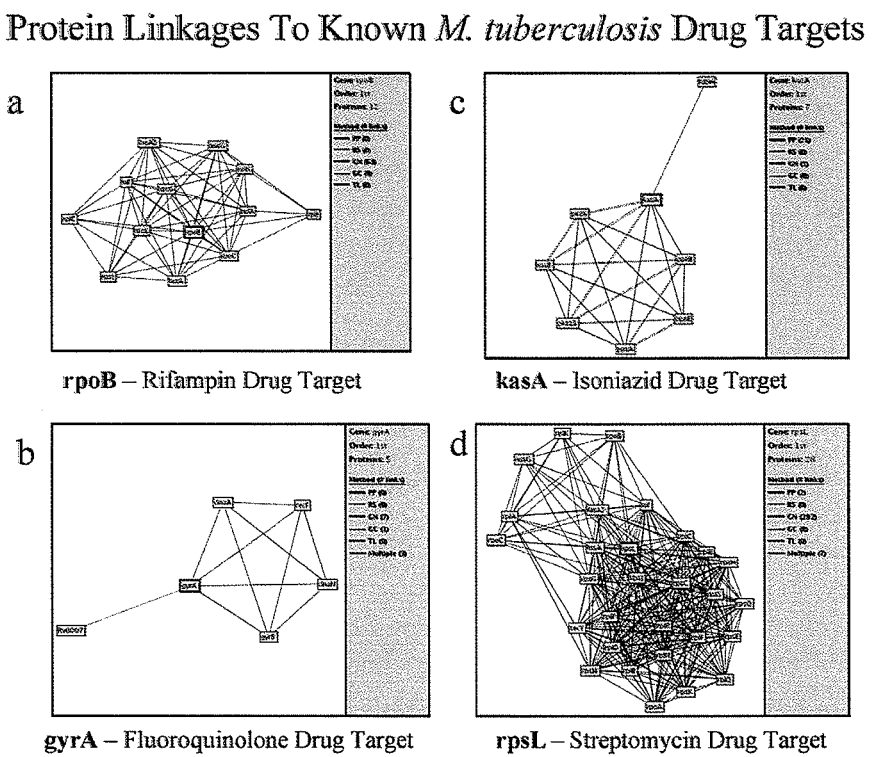


Figure 8.
Protein networks involving known *M. tuberculosis* drug targets. a) rpoB protein network (rifampin drug target), b) gyrA protein network (fluoroquinolone drug target), c) kasA protein network (isoniazid drug target), d) rpsL protein network (streptomycin drug target).

Linkage of known drug targets to uncharacterized proteins may not only suggest a potential function for these uncharacterized proteins, but may also suggest relevant leads for drug target discovery. Figures 8c and 8d show protein networks of the Isoniazid drug target, KasA, and the Streptomycin target, RpsL.

Protein networks in Figure 9 illustrate two *Streptococcus pneumoniae* drug targets, the penicillin binding proteins and the gyrase A subunit. Interestingly, the penicillin binding protein network also contains the vancomycin resistance operon member, VncR, as well as the Mur gene products, which are also involved in cell wall biosynthesis. Together, networks

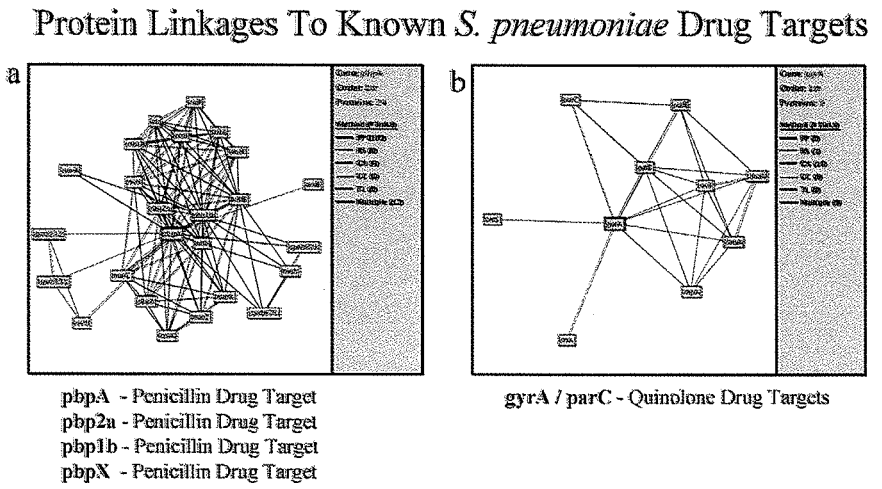


Figure 9.
Protein networks involving *S. pneumonia* drug targets. a) penicillin binding protein network, b) gyrA/parC protein network (quinolone drug targets). Protein networks such as these can be useful in identifying alternative drug targets.

such as these may suggest alternative drug targets as bacteria become resistant to current drugs.

In addition to suggesting alternative targets linked to current drug targets, protein networks can also help identify new drug targets that are associated with novel protein pathways, complexes, or cellular functions. Jeong et al. demonstrated that protein networks could be used to identify essential proteins, or proteins that are necessary for growth and survival. They found that proteins with higher connectivity in protein networks were more likely to be essential proteins, as compared to less connected proteins [58]. Essential proteins may provide useful drug targets, since the disruption of individual proteins may result in non-viable pathogens [59].

The methods described are not without noise, and methods such as the yeast two-hybrid assay are known to yield false positives in several cases. To address this situation, a number of methods have been developed to assess the reliability of various protein interaction datasets and methods for detecting protein interactions and protein linkages [60–62]. Such analyses are important, particularly when deciding which targets to pursue further.

From malaria to tuberculosis, protein networks have enabled researchers to identify and probe the global connectivity of proteins in relevant, disease-causing organisms [50, 63]. In some cases, such as in Plasmodium falciparum, protein connectivity differs from pathogenic to non-pathogenic organisms [64]. These networks enable researchers to better understand pathogens at the molecular level, and in turn can be used to identify novel drug targets. Such an approach facilitates a molecular approach to drug discovery, since drug targets are selected first at the molecular level, and then later tested at the cellular level. This is in contrast to the classical method of drug discovery, which identifies new drug compounds first at the cellular level, and later identifies the molecular target of the drug [65]. It is likely, that a combination of the two methods will yield the most promising results.

Some drugs, such as the breast cancer drug Herceptin, target the interactions between proteins. Specifically, Herceptin inhibits protein–protein interactions by binding to the extracellular domain of the human epidermal growth factor receptor, HER-2. Since protein networks often represent or suggest proteins that physically interact, protein networks may be useful for identifying relevant protein–protein interactions to target for disruption. Such a strategy is not without its challenges [66, 67], since interaction interfaces often lack amenable 'grooves' or 'binding sites' that are commonly targeted by small molecule drugs. As combinatorial drug screening advances, however, this may become an increasingly important area of focus in drug design and development.

7 Conclusion

Just as protein networks have helped us better understand the connectivity of proteins throughout the cell, protein networks also hold the promise to aid the identification of novel drug targets. As more pathogens become resistant to commonly used therapeutic agents, it will become increasingly important to pursue new strategies to combat disease. Specifically, protein networks can aid the identification of alternative protein drug targets that are linked to current drug targets, that are likely to be essential (based on network connectivity), and are linked to essential protein pathways

or complexes. Protein networks also facilitate strategies that aim to target multiple proteins of the same pathway or complex. Analysis of gene expression within the context of protein networks can also help identify proteins and protein modules that may be important for virulence. Together, protein networks can help us better understand both normal and disease mechanisms at the protein level, and in turn may provide clues to identify more effective strategies to combat disease.

Acknowledgements

The authors thank the Howard Hughes Medical Institute, National Institutes of Health, and Department of Energy-Biological and Environmental Research (DOE-BER) for support.

References

1 Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. *Nature* 405: 823–826

2 Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P et al (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403: 623–627

3 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 98: 4569–4574

4 Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141–147

5 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K et al (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180–183

6 Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21: 33–37

7 Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863–14868

8 Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285: 751–753

9 Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86–90

10 Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18 Suppl 1: S329–336

11 Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96: 4285–4288
12 Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA* 97: 6652–6657
13 Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D (2003) Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol* 4: R59
14 Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 96: 2896–2901
15 Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23: 324–328
16 Wuchty S (2002) Interaction and domain networks of yeast. *Proteomics* 2: 1715–1723
17 Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. *Nature* 407: 651–654
18 Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402: 83–86
19 Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297: 1551–1555
20 Rives AW, Galitski T (2003) Modular organization of cellular networks. *Proc Natl Acad Sci USA* 100: 1128–1133
21 Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340: 245–246
22 Vidal M, Legrain P (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res* 27: 919–929
23 Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T et al (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303: 540–543
24 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E et al (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302: 1727–1736
25 Rual JF, Venkatesan K, Hao T, Hirozane-ishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N et al (2005) Towards a proteome-scale map of the human interactome network. *Nature* 437: 1173–1178
26 [Authors?] (2005) Identification of associated proteins by coimmunoprecipitation. *Nature Methods* 2: 475–476
27 Database of Interacting Proteins: http://dip.doe-mbi.ucla.edu/
28 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32: D449–451
29 Gilbert D (2005) Biomolecular interaction network database. *Brief Bioinform* 6: 194–198

30 Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a Molecular INTeraction database. *FEBS Lett* 513: 135–140

31 Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* 30: 42–46

32 Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res* 33: D334–337

33 Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res* 34: D169–172

34 Pellegrini M, Thompson M, Fierro J, Bowers P (2001) Computational method to assign microbial genes to pathways. *J Cell Biochem* 37: 106–109

35 GOLD – Genomes OnLine Database. http://www.genomesonline.org/

36 Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* 29: 126–127

37 Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE et al (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537–544

38 Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 5: R35

39 Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res* 29: 72–74

40 ProLinks Database – http://www.doe-mbi.ucla.edu/

41 von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31: 258–261

42 Hollstein M, Sidransky D, Vogelstein B, Harris CC (1991) p53 mutations in human cancers. *Science* 253: 49–53

43 Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, Filipovic Z, Kong N, Kammlott U, Lukacs C, Klein C et al (2004) *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* 303: 844–848

44 Patino MM, Liu JJ, Glover JR, Lindquist S (1996) Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science* 273: 622-626

45 Serio TR, Cashikar AG, Kowal AS, Sawicki GJ, Moslehi JJ, Serpell L, Arnsdorf MF, Lindquist SL (2000) Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* 289: 1317-1321

46 DePace AH, Santoso A, Hillner P, Weissman JS (1998) A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion. *Cell* 93: 1241-1252

47 Balbirnie M, Grothe R, Eisenberg DS (2001) An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc Natl Acad Sci USA* 98: 2375-2380

48 Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, Grothe R, Eisenberg D (2005) Structure of the cross-beta spine of amyloid-like fibrils. *Nature* 435: 747–749

49 Shorter J, Lindquist S (2004) Hsp104 catalyzes formation and elimination of self-replicating Sup35 prion conformers. *Science* 304: 1793–1797

50 Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (2003) Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Res* 31: 7099–7109

51 Yeats C, Finn RD, Bateman A (2002) The PASTA domain: a beta-lactam-binding domain. *Trends Biochem Sci* 27: 438

52 Rachman H, Strong M, Schaible U, Schuchhardt J, Hagens K, Mollenkopf H, Eisenberg D, Kaufmann SHE (2006) *Mycobacterium tuberculosis* gene expression profiling within the context of protein networks. *Microbes and Infection* 8: 747–757

53 D'Costa VM, McGrann KM, Hughes DW, Wright GD (2006) Sampling the antibiotic resistome. *Science* 311: 342–343

54 Blower SM, Chou T (2004) Modeling the emergence of the 'hot zones': tuberculosis and the amplification dynamics of drug resistance. *Nature Medicine* 10: 1111–1116

55 Hecht FM, Grant RM, Petropoulos CJ, Dillon B, Chesney MA, Tian H, Hellmann NS, Bandrapalli NI, Digilio L, Branson B et al (1998) Sexual transmission of an HIV-1 variant resistant to multiple reverse-transcriptase and protease inhibitors. *N Engl J Med* 339: 307–311

56 Ohsaki Y, Tachibana M, Nakanishi K, Nakao S, Saito K, Toyoshima E, Sato M, Takahashi T, Osanai S, Itoh Y et al (2003) Alterations in penicillin binding protein gene of *Streptococcus pneumoniae* and their correlation with susceptibility patterns. *Int J Antimicrob Agents* 22: 140–146

57 Hatful G, Jacobs WR (eds) (2000) *Molecular Genetics of Mycobacteria*. American Society Microbiology Press, New York, USA

58 Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411: 41–42

59 Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 48: 77–84

60 Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteom* 1: 349–356

61 von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417: 399–403

62 Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein–protein interaction data? *J Mol Biol* 327: 919–923

63 LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C et al (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438: 103–107

64 Suthram S, Sittler T, Ideker T (2005) The Plasmodium protein network diverges from those of other eukaryotes. *Nature* 438: 108–112

65 Chaudhuri A, Chant J (2005) Protein-interaction mapping in search of effective drug targets. *Bioessays* 27: 958–969

66 Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat Rev Drug Discov* 3: 301–317

67 Zhao L, Chmielewski J (2005) Inhibiting protein–protein interactions using designed molecules. *Curr Opin Struct Biol* 15: 31–34

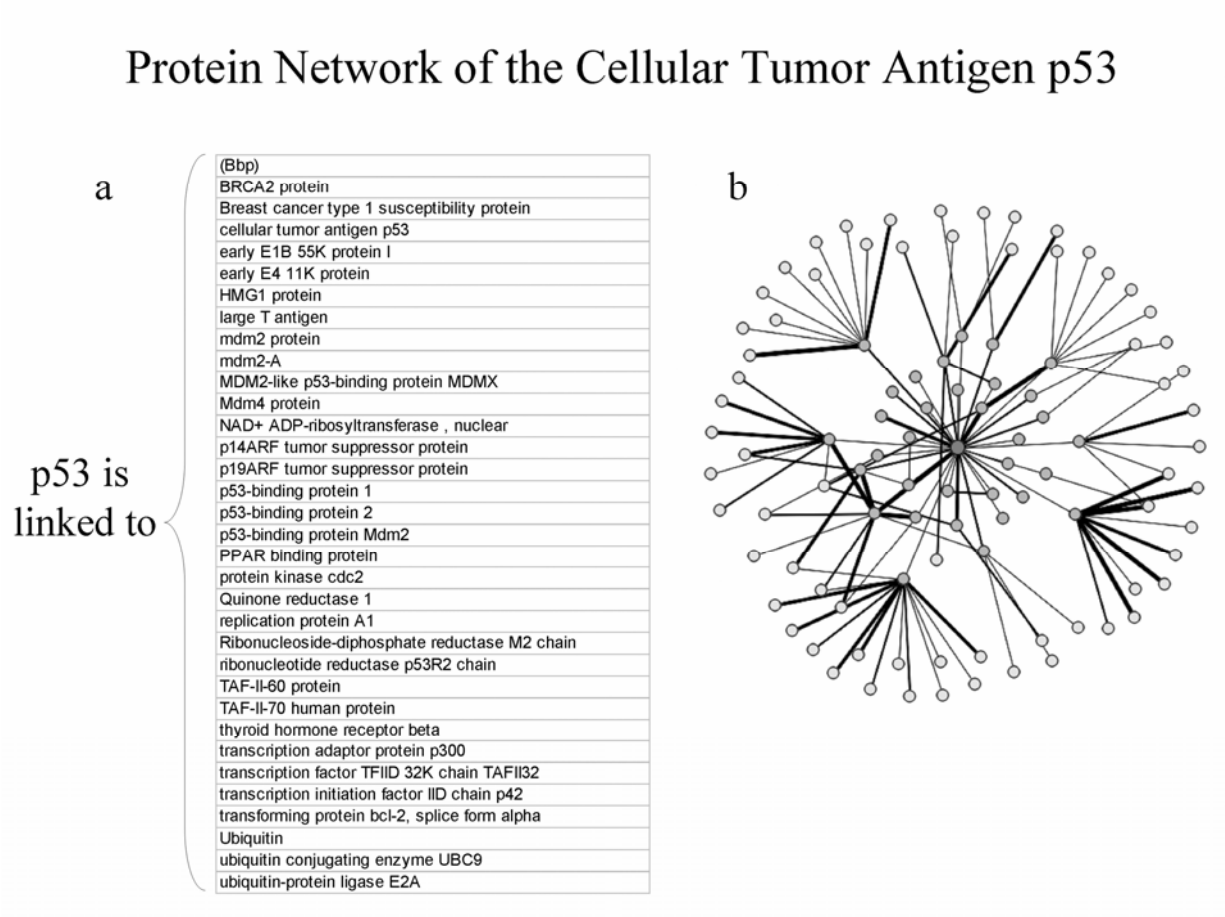
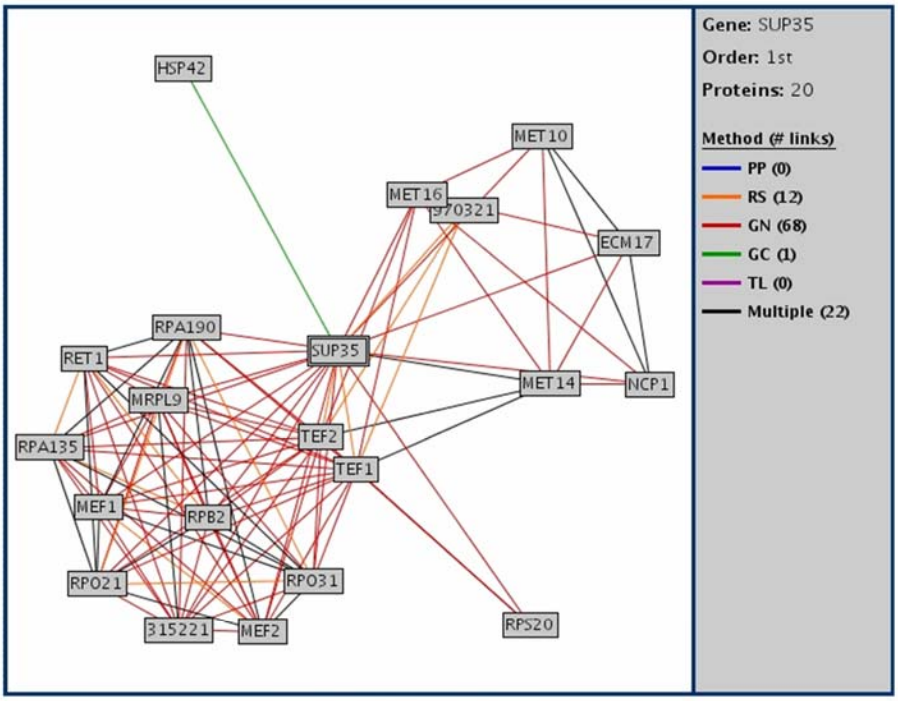

Protein Network of the Cellular Tumor Antigen p53
a
p53 is linked to
(Bbp)
BRCA2 protein
Breast cancer type 1 susceptibility protein
cellular tumor antigen p53
early E1B 55K protein I
early E4 11K protein
HMG1 protein
large T antigen
mdm2 protein
mdm2-A
MDM2-like p53-binding protein MDMX
Mdm4 protein
NAD+ ADP-ribosyltransferase , nuclear
p14ARF tumor suppressor protein
p19ARF tumor suppressor protein
p53-binding protein 1
p53-binding protein 2
p53-binding protein Mdm2
PPAR binding protein
protein kinase cdc2
Quinone reductase 1
replication protein A1
Ribonucleoside-diphosphate reductase M2 chain
ribonucleotide reductase p53R2 chain
TAF-II-60 protein
TAF-II-70 human protein
thyroid hormone receptor beta
transcription adaptor protein p300
transcription factor TFIID 32K chain TAFII32
transcription initiation factor IID chain p42
transforming protein bcl-2, splice form alpha
Ubiquitin
ubiquitin conjugating enzyme UBC9
ubiquitin-protein ligase E2A
b

Figure 1

Protein Network of the Yeast Prion Protein Sup35



Sup35 -Yeast Prion Protein

Figure 2

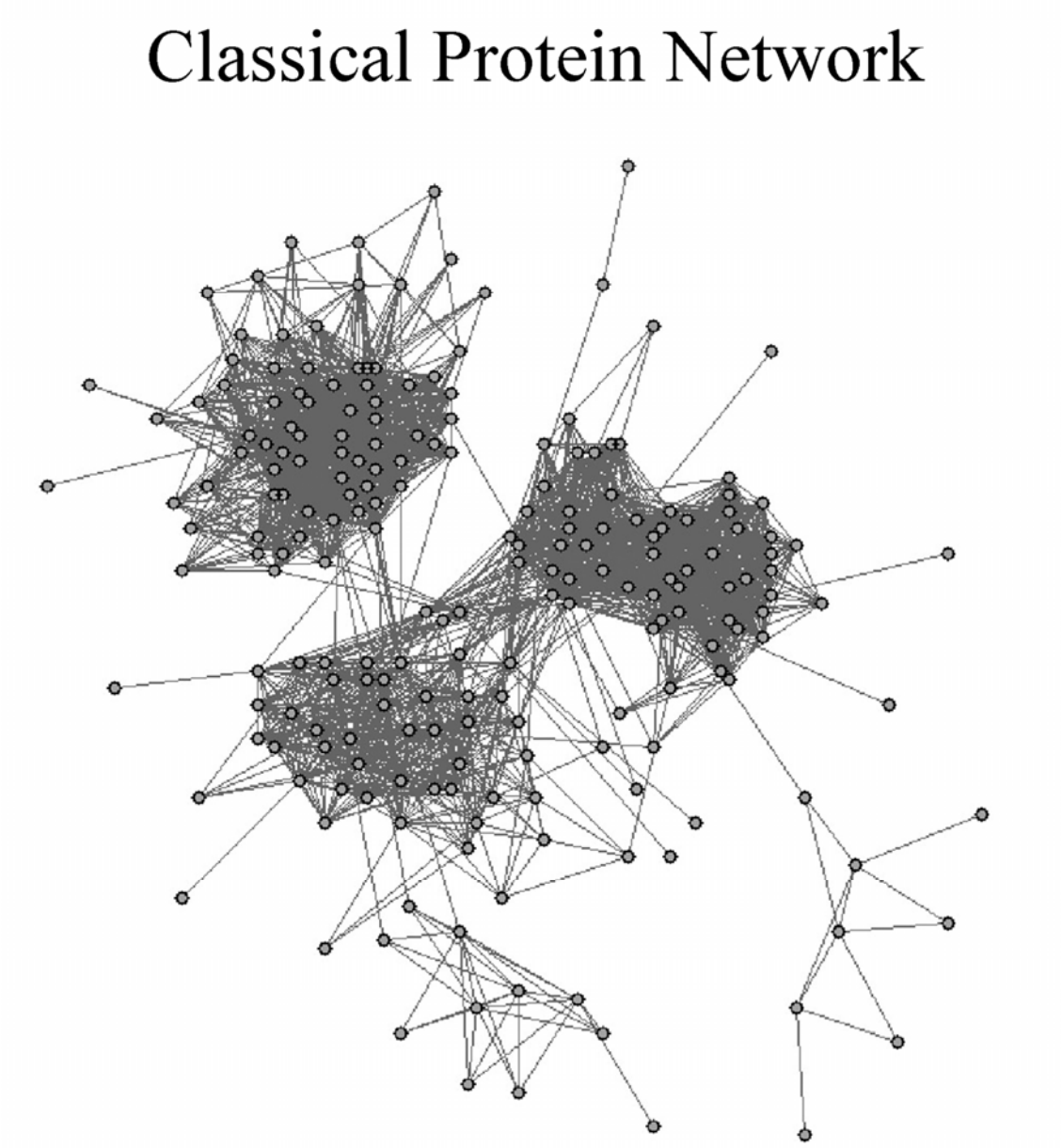


Figure 3

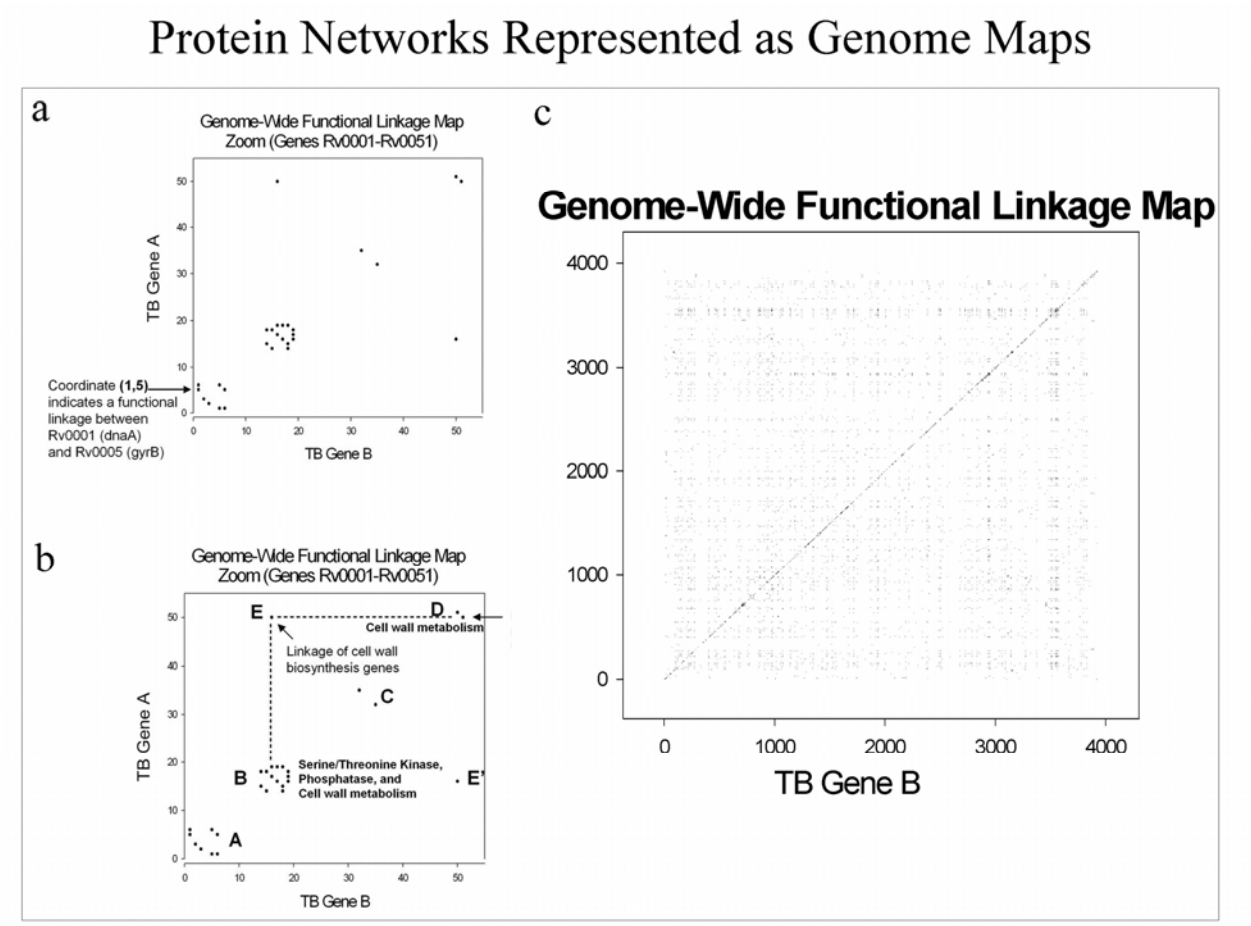

Protein Networks Represented as Genome Maps
a
Genome-Wide Functional Linkage Map
Zoom (Genes Rv0001-Rv0051)
TB Gene A
TB Gene B
Coordinate (1,5) indicates a functional linkage between Rv0001 (dnaA) and Rv0005 (gyrB)
b
Genome-Wide Functional Linkage Map
Zoom (Genes Rv0001-Rv0051)
E
D
Cell wall metabolism
Linkage of cell wall biosynthesis genes
C
B
Serine/Threonine Kinase, Phosphatase, and Cell wall metabolism
E'
A
TB Gene A
TB Gene B
c
Genome-Wide Functional Linkage Map
TB Gene B

Figure 4

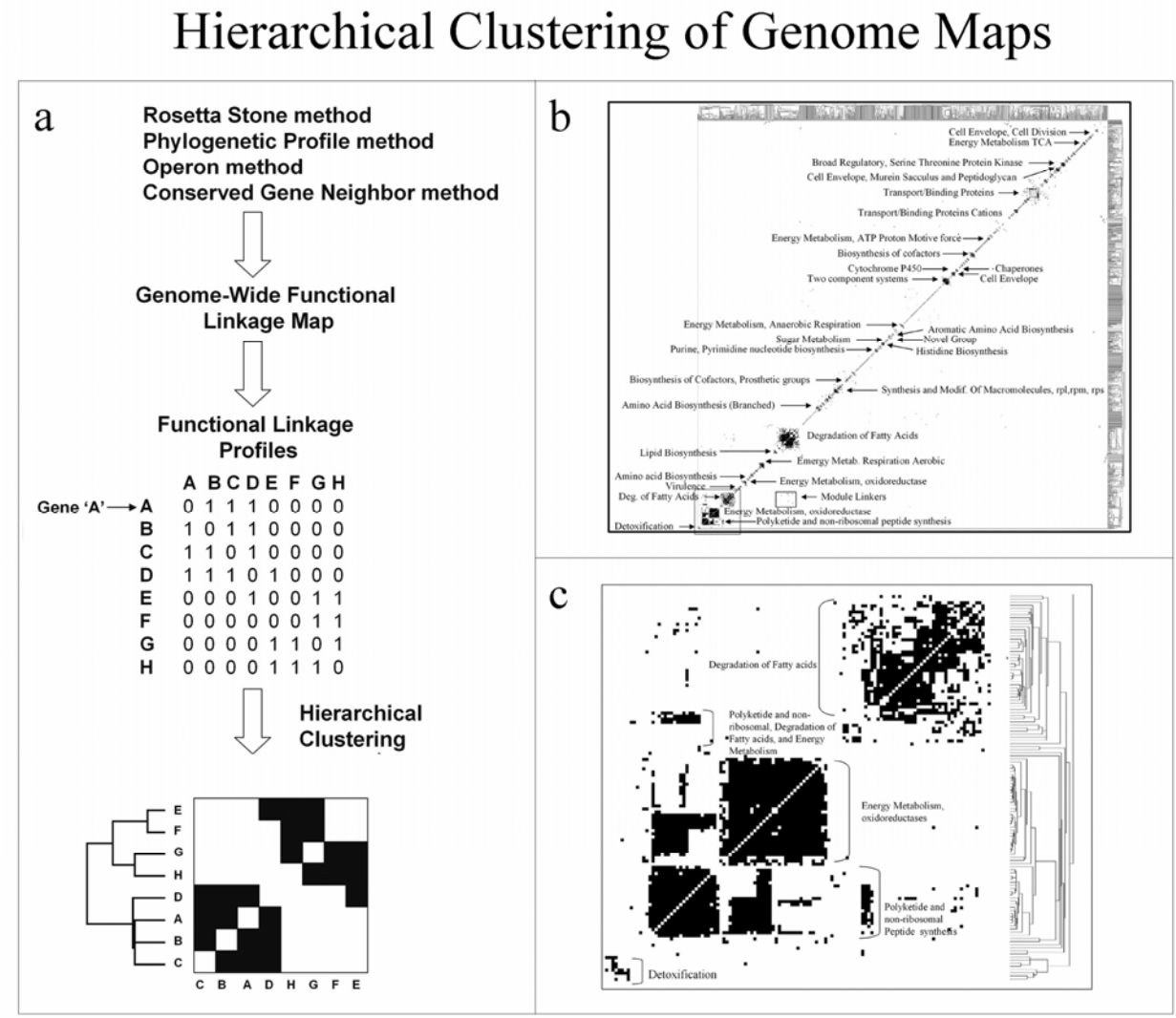

Hierarchical Clustering of Genome Maps
a
Rosetta Stone method
Phylogenetic Profile method
Operon method
Conserved Gene Neighbor method
Genome-Wide Functional Linkage Map
Functional Linkage Profiles
Gene 'A'
A B C D E F G H
A 0 1 1 1 0 0 0 0
B 1 0 1 1 0 0 0 0
C 1 1 0 1 0 0 0 0
D 1 1 1 0 1 0 0 0
E 0 0 0 1 0 0 1 1
F 0 0 0 0 0 0 1 1
G 0 0 0 0 1 1 0 1
H 0 0 0 0 1 1 1 0
Hierarchical Clustering
E F G H D A B C
C B A D H G F E
b
Cell Envelope, Cell Division
Energy Metabolism TCA
Broad Regulatory, Serine Threonine Protein Kinase
Cell Envelope, Murein Sacculus and Peptidoglycan
Transport/Binding Proteins
Transport/Binding Proteins Cations
Energy Metabolism, ATP Proton Motive force
Biosynthesis of cofactors
Cytochrome P450
Two component systems
Chaperones
Cell Envelope
Energy Metabolism, Anaerobic Respiration
Sugar Metabolism
Purine, Pyrimidine nucleotide biosynthesis
Aromatic Amino Acid Biosynthesis
Novel Group
Histidine Biosynthesis
Biosynthesis of Cofactors, Prosthetic groups
Amino Acid Biosynthesis (Branched)
Synthesis and Modif. Of Macromolecules, rpl,rpm, rps
Degradation of Fatty Acids
Lipid Biosynthesis
Emergy Metab. Respiration Aerobic
Amino acid Biosynthesis
Virulence
Energy Metabolism, oxidoreductase
Deg. of Fatty Acids
Module Linkers
Energy Metabolism, oxidoreductase
Detoxification
Polyketide and non-ribosomal peptide synthesis
c
Degradation of Fatty acids
Polyketide and non-ribosomal, Degradation of Fatty acids, and Energy Metabolism
Energy Metabolism, oxidoreductases
Polyketide and non-ribosomal Peptide synthesis
Detoxification

Figure 5

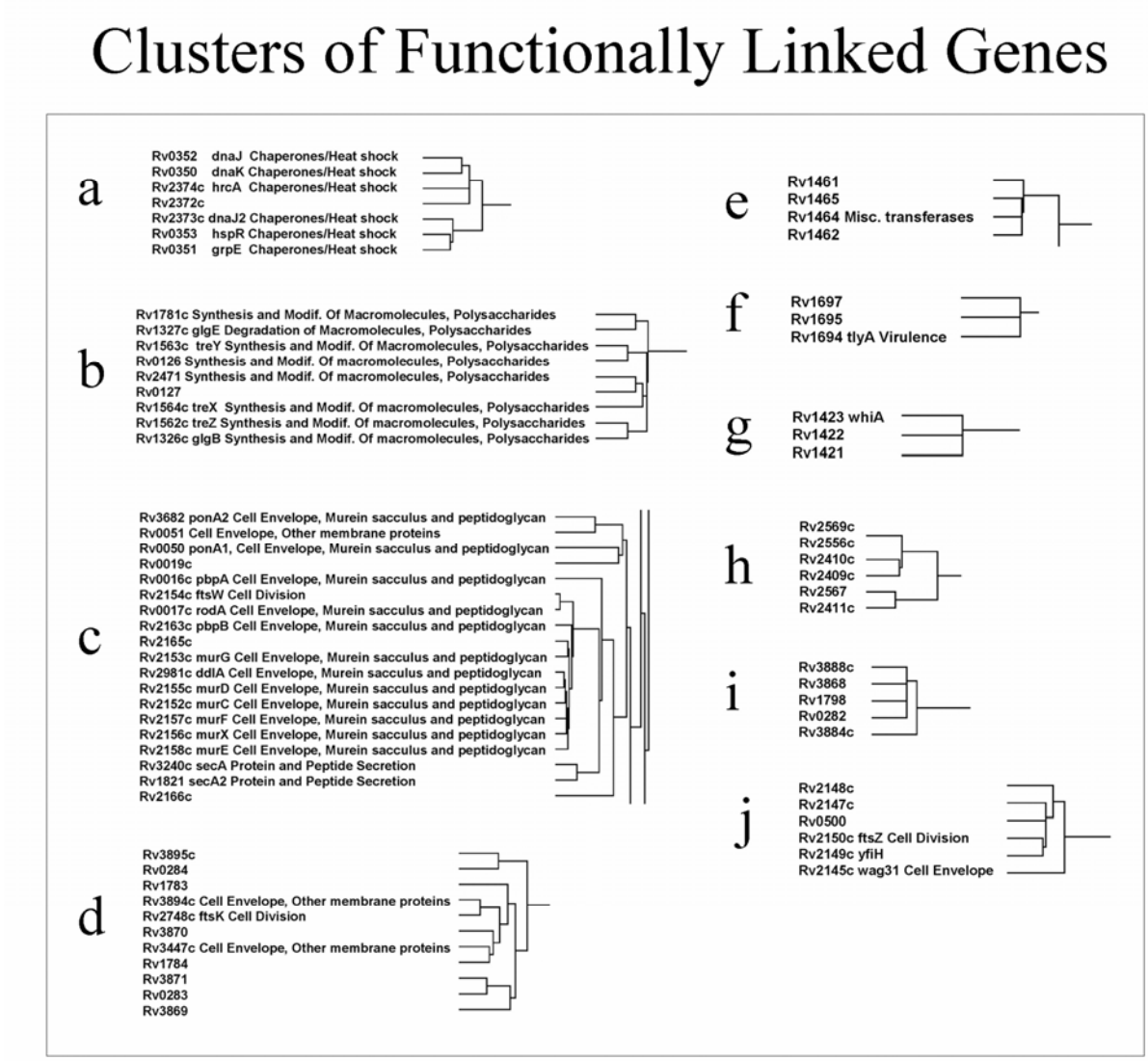

Clusters of Functionally Linked Genes
a
Rv0352 dnaJ Chaperones/Heat shock
Rv0350 dnaK Chaperones/Heat shock
Rv2374c hrcA Chaperones/Heat shock
Rv2372c
Rv2373c dnaJ2 Chaperones/Heat shock
Rv0353 hspR Chaperones/Heat shock
Rv0351 grpE Chaperones/Heat shock
b
Rv1781c Synthesis and Modif. Of Macromolecules, Polysaccharides
Rv1327c glgE Degradation of Macromolecules, Polysaccharides
Rv1563c treY Synthesis and Modif. Of Macromolecules, Polysaccharides
Rv0126 Synthesis and Modif. Of macromolecules, Polysaccharides
Rv2471 Synthesis and Modif. Of macromolecules, Polysaccharides
Rv0127
Rv1564c treX Synthesis and Modif. Of macromolecules, Polysaccharides
Rv1562c treZ Synthesis and Modif. Of macromolecules, Polysaccharides
Rv1326c glgB Synthesis and Modif. Of macromolecules, Polysaccharides
c
Rv3682 ponA2 Cell Envelope, Murein sacculus and peptidoglycan
Rv0051 Cell Envelope, Other membrane proteins
Rv0050 ponA1, Cell Envelope, Murein sacculus and peptidoglycan
Rv0019c
Rv0016c pbpA Cell Envelope, Murein sacculus and peptidoglycan
Rv2154c ftsW Cell Division
Rv0017c rodA Cell Envelope, Murein sacculus and peptidoglycan
Rv2163c pbpB Cell Envelope, Murein sacculus and peptidoglycan
Rv2165c
Rv2153c murG Cell Envelope, Murein sacculus and peptidoglycan
Rv2981c ddlA Cell Envelope, Murein sacculus and peptidoglycan
Rv2155c murD Cell Envelope, Murein sacculus and peptidoglycan
Rv2152c murC Cell Envelope, Murein sacculus and peptidoglycan
Rv2157c murF Cell Envelope, Murein sacculus and peptidoglycan
Rv2156c murX Cell Envelope, Murein sacculus and peptidoglycan
Rv2158c murE Cell Envelope, Murein sacculus and peptidoglycan
Rv3240c secA Protein and Peptide Secretion
Rv1821 secA2 Protein and Peptide Secretion
Rv2166c
d
Rv3895c
Rv0284
Rv1783
Rv3894c Cell Envelope, Other membrane proteins
Rv2748c ftsK Cell Division
Rv3870
Rv3447c Cell Envelope, Other membrane proteins
Rv1784
Rv3871
Rv0283
Rv3869
e
Rv1461
Rv1465
Rv1464 Misc. transferases
Rv1462
f
Rv1697
Rv1695
Rv1694 tlyA Virulence
g
Rv1423 whiA
Rv1422
Rv1421
h
Rv2569c
Rv2556c
Rv2410c
Rv2409c
Rv2567
Rv2411c
i
Rv3888c
Rv3868
Rv1798
Rv0282
Rv3884c
j
Rv2148c
Rv2147c
Rv0500
Rv2150c ftsZ Cell Division
Rv2149c yfiH
Rv2145c wag31 Cell Envelope

Figure 6

Protein Networks and Gene Expression

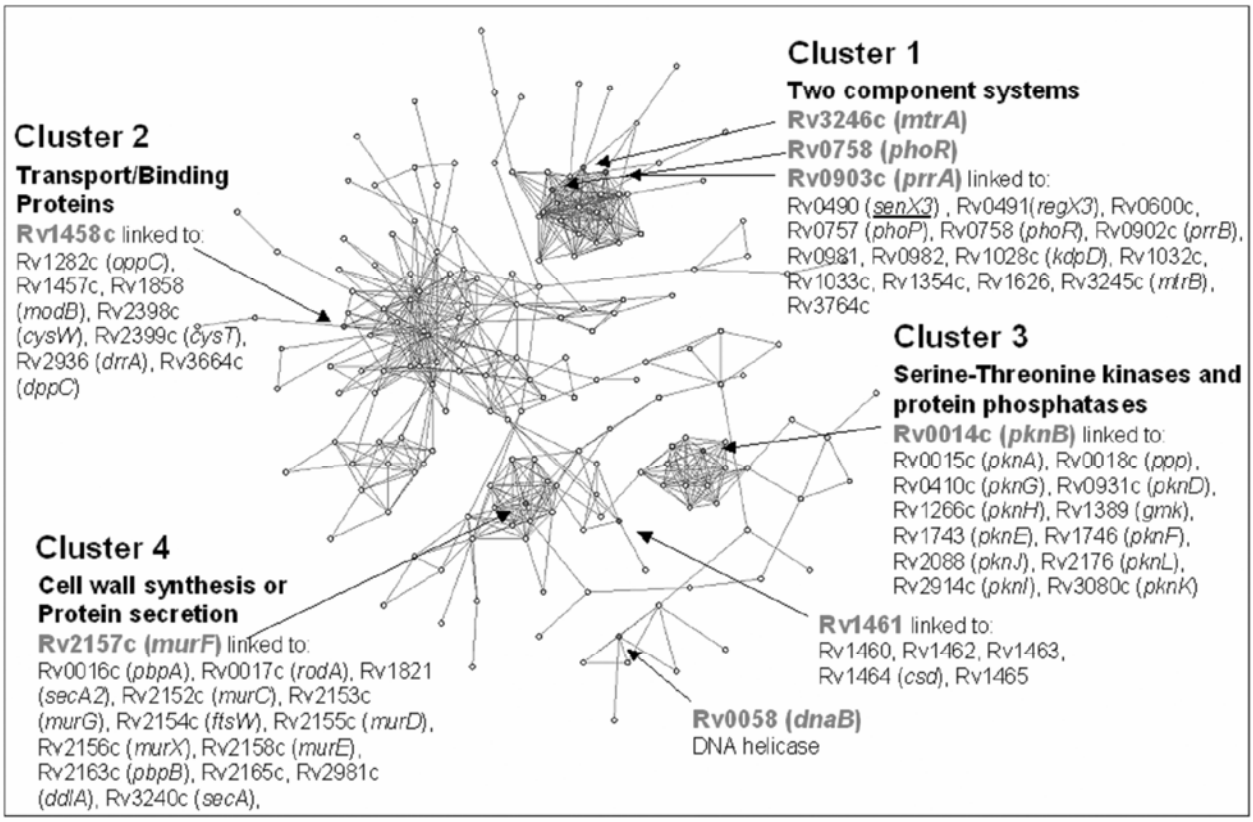


Figure 7

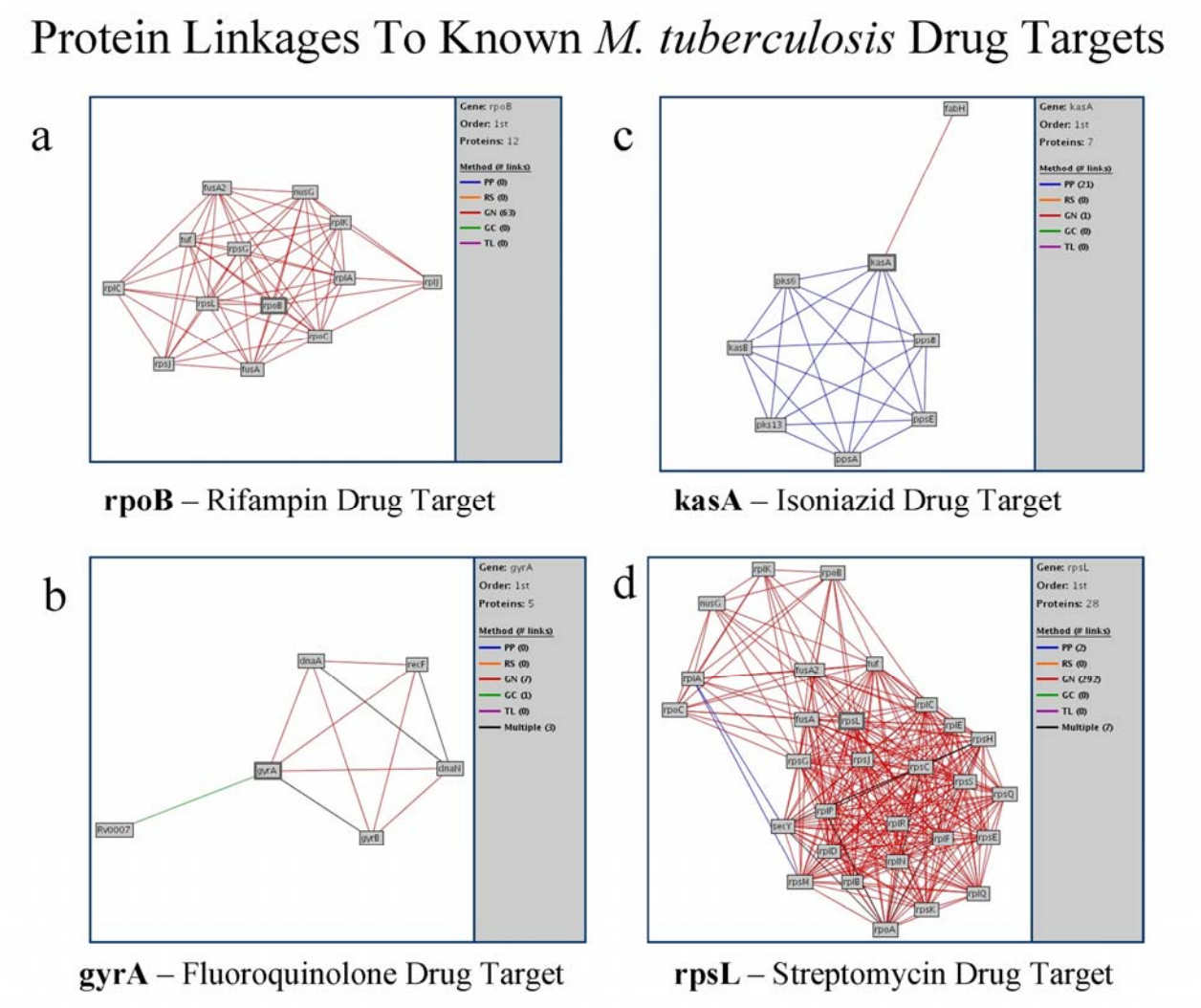
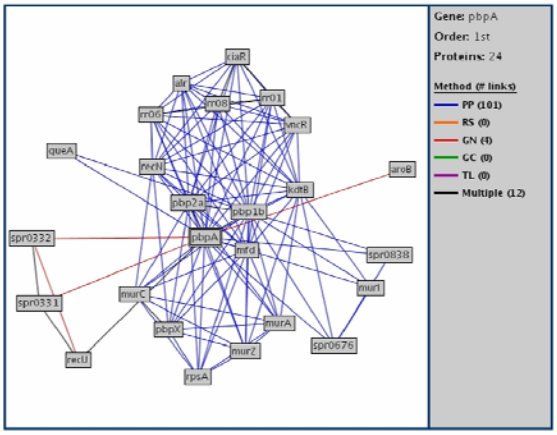

Protein Linkages To Known M. tuberculosis Drug Targets
a
Gene: rpoB
Order: 1st
Proteins: 12
Method (# links)
PP (0)
RS (0)
GN (63)
GC (0)
TL (0)
fusA2
nusG
rplK
fuf
rpsG
rplA
rplJ
rplC
rpsL
rpoB
rpoC
rpsJ
fusA
rpoB – Rifampin Drug Target
c
Gene: kasA
Order: 1st
Proteins: 7
Method (# links)
PP (21)
RS (0)
GN (1)
GC (0)
TL (0)
fabH
kasA
pks6
kasB
ppsB
pks13
ppsE
ppsA
kasA – Isoniazid Drug Target
b
Gene: gyrA
Order: 1st
Proteins: 5
Method (# links)
PP (0)
RS (0)
GN (7)
GC (1)
TL (0)
Multiple (3)
dnaA
recF
gyrA
dnaN
Rv0007
gyrB
gyrA – Fluoroquinolone Drug Target
d
Gene: rpsL
Order: 1st
Proteins: 28
Method (# links)
PP (2)
RS (0)
GN (29.7)
GC (0)
TL (0)
Multiple (7)
rplK
rpoB
nusG
rplA
fusA2
fuf
rpoC
rplC
fusA
rpsL
rplE
rpsH
rpsG
rpsJ
rpsC
rpsS
rpsQ
rplF
secY
rplR
rplF
rpsE
rplD
rplN
rpsM
rplB
rplQ
rpsK
rpoA
rpsL – Streptomycin Drug Target

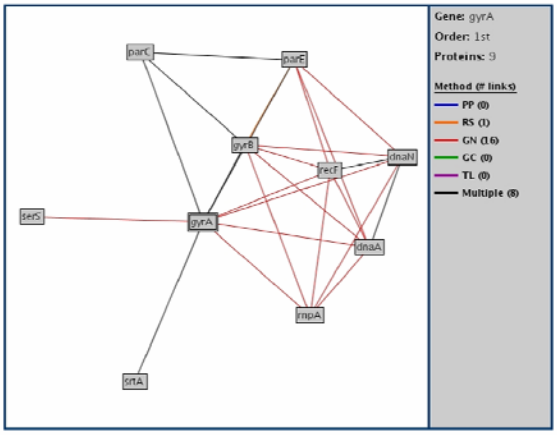
Figure 8

Protein Linkages To Known *S. pneumoniae* Drug Targets

a



b



pbpA - Penicillin Drug Target
pbp2a - Penicillin Drug Target
pbp1b - Penicillin Drug Target
pbpX - Penicillin Drug Target

gyrA / parC - Quinolone Drug Targets

Figure 9