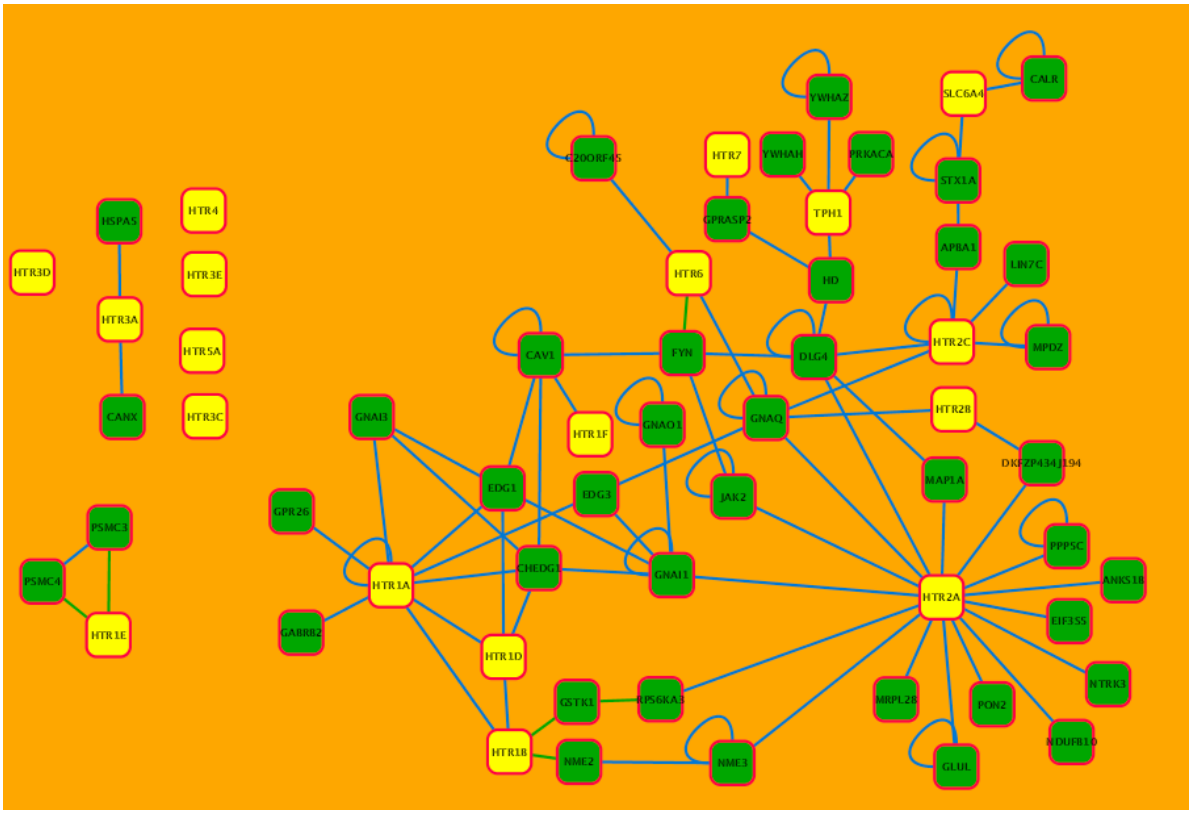
Ramón Aragüés Peleato

Protein interaction networks and their applications to protein characterization and cancer genes prediction



PhD Thesis

Barcelona, May 2007

The image of the cover shows the happiness protein interaction network (i.e. the protein interaction network for proteins involved in the serotonin pathway)

Protein Interaction Networks and their Applications to Protein Characterization and Cancer Genes Prediction

Ramón Aragüés Peleato

Memòria presentada per optar al grau de
Doctor en Biologia per la Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del Dr. Baldo Oliva al Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra

Baldo Oliva Miguel Ramón Aragüés Peleato

Barcelona, Maig 2007

The research in this thesis has been carried out at the Structural Bioinformatics Lab (SBI) within the Grup de Recerca en Informàtica Biomèdica at the Parc de Recerca Biomèdica de Barcelona (PRBB).

The research carried out in this thesis has been supported by a “Formación de Personal Investigador (FPI)” grant from the Ministerio de Educación y Ciencia awarded to Dr. Baldo Oliva.

A mis padres, que me hicieron querer conseguirlo
A Natalia, que me ayudó a conseguir hacerlo

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I know this is the section of the thesis that most people read; in fact, it might be the only section of the thesis most people ever think about reading. Maybe in your case this acknowledgements thing is even more pronounced than it usually is, and it has become the central part of your life and it is all you care about. Or perhaps, you didn't even know there was an acknowledgements section in a PhD thesis, and you just told yourself: "let's read it!". Whatever it is that made you read this, you are going to regret it, because these are going to be the most disappointing acknowledgements in human history. At least, for those who do not know the answers…

http://www.aragues.com/thesis_ack.php

THESIS ABSTRACT

The importance of understanding cellular processes has prompted the development of experimental approaches that detect protein-protein interactions. Recently, most approaches have been focused on large-scale screenings of protein-protein interactions, such as two-hybrid assays and affinity purifications followed by mass spectrometry. Computational methods can use protein interaction data for tasks such as protein annotation and protein interactions prediction. However, due to data spread and disparate storage formats, protein interactions research has been subjected to using a subset of all information available.

Here, we describe a software platform called PIANA (Protein Interactions And Network Analysis) that facilitates working with protein interactions by integrating data from multiple sources and automating the analysis of protein interaction networks. For example, PIANA can be used to retrieve all interactions of a given protein, create the interaction network of a particular disease, transfer protein interactions from model organisms to human, and to map gene expression information into a protein interaction network.

Experimental methods for protein interactions detection do not identify the protein interfaces involved in interactions. Here, we describe a method implemented within PIANA for delineating the interacting motifs of proteins. We rely on the observation that proteins with common interaction partners tend to interact with these partners through a common interacting motif. The positive predictive value of our method in detecting proteins with common SCOP families is 75% at sensitivity of 10%. We find that highly connected proteins in the network (i.e., hubs) with multiple interacting motifs are more likely to be essential than hubs with one or two interacting motifs, thus rationalizing the previously observed correlation between essentiality and the number of interacting partners of a protein.

Cancer is a complex disease, involving multiple and specific changes at the DNA level that can be inherited or induced by environmental factors. Data from genomics and proteomics projects can be used to identify proteins involved in cancer. Here, we present a method that predicts cancer genes by integrating protein-protein interaction data, differential expression studies and structural, functional and evolutionary properties. For a minimum sensitivity of 1%, our approach obtained a positive predictive value of 71%, which is higher than the positive predictive value achieved by any of the methods independently.

CHAPTER I

INTRODUCTION

This thesis is divided into chapters, and each chapter contains one or more articles that have been published (or submitted) in journals rated in Journal Citation Reports. At the beginning of each chapter, we provide a non-too-technical short overview of the chapter, for those wishing to get a grasp of the work presented in this thesis (but do not wish to plunge into its contents).

This introductory chapter starts by providing an overview of molecular biology and summarizing the experimental techniques that are being used to extract meaningful data from biological systems. The rest of the chapter is dedicated to describing protein interaction networks and their applications, together with a brief state-of-the-art on other related areas.

The thesis ends with an epilogue where the author tries to explain using too many words something that Matt Cartmill summarized in much shorter terms:

"As an adolescent I aspired to lasting fame, I craved factual certainty, and I thirsted for a meaningful vision of human life - so I became a scientist. This is like becoming an archbishop so you can meet girls."

1.1 Introduction overview

The completion of genome sequencing projects stimulated the development of high-throughput experimental methods aimed at functional characterization of the discovered genes. In particular, the identification of protein-protein interactions has been accelerated by the development of new technologies. Thus, a vast amount of protein-protein interaction data has been collected, including proteome-scale interactome maps for yeast [1, 2], fly [3] and worm [4], and a partial map for human [5, 6]. The analysis of these maps has shown potential for providing insights about biological systems [7, 8]. However, interaction data is spread across multiple repositories, which hinders the access to all known information. The objective of this thesis was to contribute towards the optimal integration of all available protein-protein interaction data and the use of this data for providing biological insights about proteins, protein interactions, and protein function (or dysfunction, such as in cancer).

This introductory chapter shortly describes the biological mechanisms related to protein-protein interactions and provides the general bioinformatics background needed for understanding the key contributions of the work presented here. More specific descriptions of the state of the art can be found in the introductions of the articles included in this thesis.

1.2 Background

1.2.1 Molecular cell biology

Deoxyribonucleic acid (DNA) is the cellular library that contains all the information required to build the cells and tissues of an organism. The exact duplication of this information in any species from generation to generation assures the genetic continuity of that species. The information is arranged in hereditary units (i.e. genes) that control the identifiable traits of an organism. In the process of transcription (Figure 1.1), the information stored in DNA is copied into ribonucleic acid (RNA), which has three distinct roles in protein synthesis. Messenger RNA (mRNA) carries the instructions from DNA that specify the order of amino acids during protein synthesis. The assembly of amino acids into a protein occurs by translation of mRNA (Figure 1.1). In this process, the information in mRNA is interpreted by a second type of RNA called transfer RNA (tRNA) with the aid of a third type of RNA, ribosomal RNA (rRNA), and its associated proteins. As the correct amino acids are brought into sequence by tRNAs, they are linked by peptide bonds to make proteins [9].

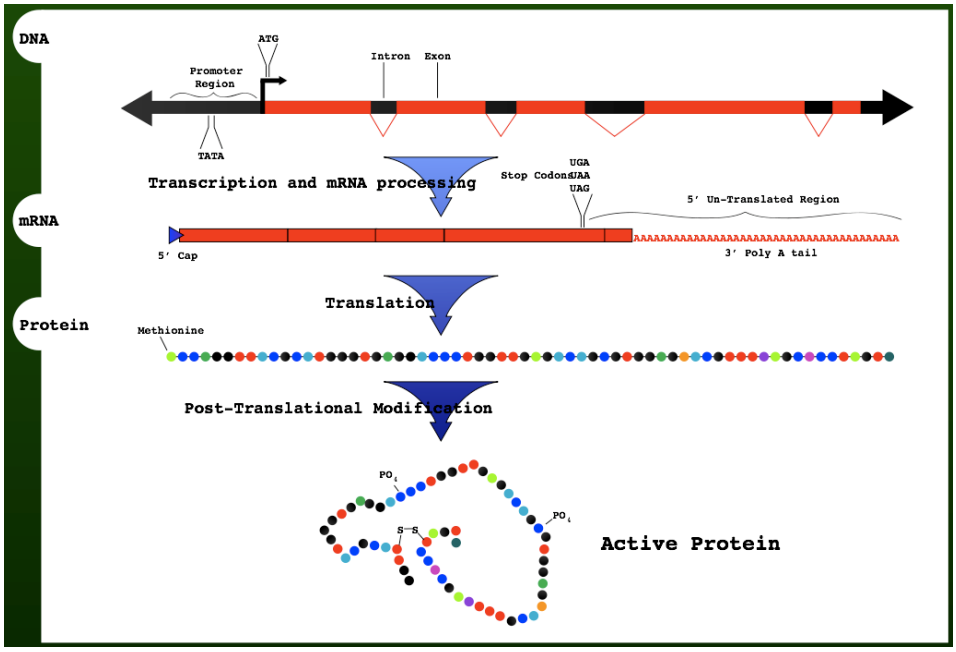


Figure 1.1 From DNA to mRNA to protein. *(Image by Mike Jones, licensed under the Creative Commons Attribution ShareAlike License v. 2.5)*

The process by which a gene is "turned on" to produce the specific biological molecule encoded by that gene (usually protein or RNA) is referred as "gene expression". The process by which the cell controls when and where genes will be activated and how much gene product will be produced is called "gene regulation". Gene regulation is usually achieved through interactions among DNA, RNA and proteins.

Gene products are the active agents of the cell. In particular, proteins are said to be the chief actors of the cell, responsible for carrying out the program of activities encoded by genes[2] [9]. The main property that enables proteins to carry out their diverse cellular functions is their ability to bind (i.e. to interact with) other molecules, either in permanent complexes or in transient interactions. The protein surface that is physically in contact with the other molecule during an interaction is referred as protein interface. Proteins functions range from providing structural support to the cell to acting as enzymes that promote specific chemical reactions or acting as a transcription factor (i.e. binding to DNA to regulate transcription). In order to perform these functions, most proteins must first fold into a three-dimensional (3D) structure, which gives them their specific chemical functionality

[2] The word "protein" derives from the Greek word "protos" meaning first.

(Figure 1.2). There are four different aspects to the structure of a protein: (i) primary structure, referring to the amino acid sequence of the protein; (ii) secondary structure, referring to the general 3D form of local protein segments; (iii) tertiary structure, referring to the 3D structure of a single protein molecule; and (iv) cuaternary structure, referring to complexes of several protein (or other) molecules. The most common secondary structures are alpha helices and beta sheets. The tertiary structure of a protein is usually built from one or more domains, which are protein subunits capable of folding autonomously from the rest of the protein. Usually, a domain has a functionality of its own, and the same type of domain can be found in multiple –and in many occasions, unrelated- proteins.

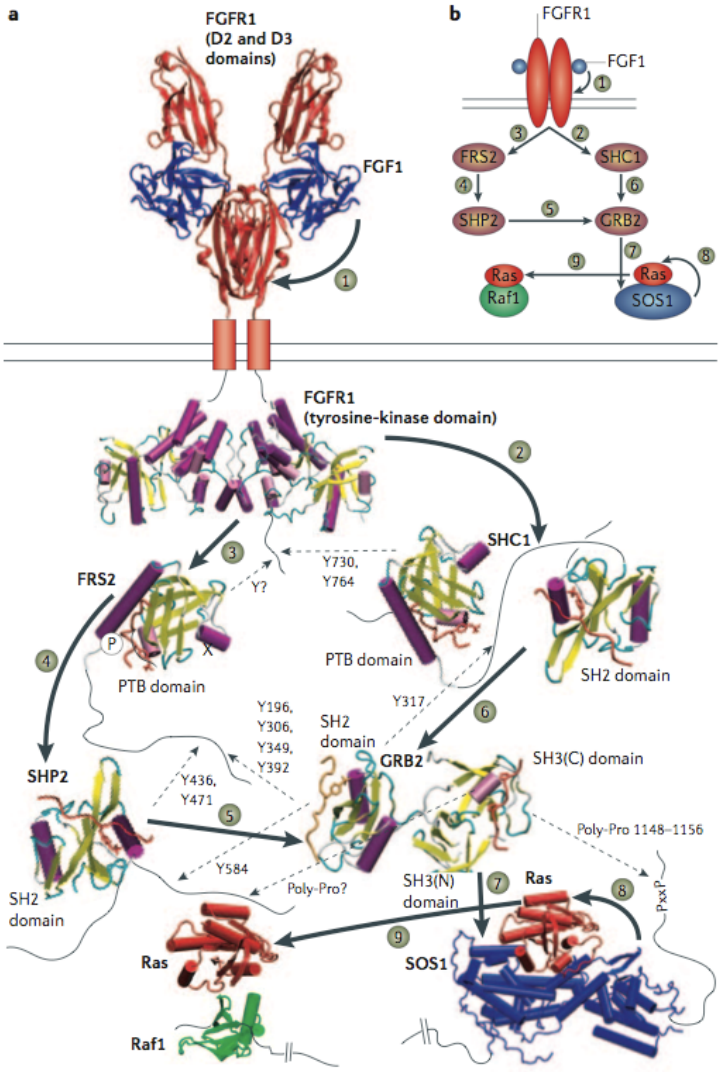


Figure 1.2 Part of one biological pathway (fibroblast growth factor signaling) from a structural perspective. Image obtained from [10]

1.2.2 Bioinformatics and Computational Biology

According to the USA National Institute of Health (NIH), bioinformatics and computational biology are defined [11, 12] as follows:

- *Bioinformatics*: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- *Computational Biology*: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Along this thesis, we will use the terms *bioinformatics* and *computational biology* indistinctively to refer to the development and application of computational tools (including theoretical methods and mathematical models) for studying biological systems and data.

1.3 Genomics and proteomics

The term 'genome' is used to refer to the complete genetic content of an organism. Analogously, the term 'proteome' is used to refer to the entire protein complement of a given genome, that is, the complete set of proteins made by a given organism in a given cell at a particular point in time[3] [13]. In the last decade, one-by-one study of genes and proteins has been replaced by 'genomics' and 'proteomics', in which scientists attempt to study all genes or proteins comprehensively and simultaneously. Both genomics and proteomics have produce a plethora of data that needs to be systematically organized and analyzed in order to make significant contributions to our understanding of the biology of the cell. In this section, we briefly describe some[4] experimental techniques designed to perform high-throughput extraction of genomic and proteomic data.

1.3.1 Two-dimensional gel electrophoresis and mass spectrometry

In order to identify proteins of interest in a given tissue or cell, researchers usually combine two-dimensional gel electrophoresis with mass spectrometry [14]. In this approach, labeled proteins from a cell or tissue extract are separated on the gel and then analyzed by mass spectrometry, usually after digesting the proteins to produce unique degradation products.

[3] The complete proteome for an organism can be conceptualized as the complete set of proteins from all of the various cellular proteomes.
[4] This thesis has been mainly focused on protein-protein interactions. Therefore, we present here methods for the detection of protein-protein interactions and other experimental techniques that produce data related to the study of protein function.

The separation on the gel is performed on two dimensions: one on the basis of charge (i.e. isoelectric point), and another on the basis of molecular weight. This combination of techniques has provided biologist with a tool capable of resolving proteins in high-throughput mode [15].

1.3.2 Three-dimensional structure determination

The three-dimensional (3D) structure of proteins is usually determined using x-ray crystallography or nuclear magnetic resonance (NMR) [16-18]. Currently, these two techniques are being used in structural genomics projects to determine the 3D structure of as many proteins as possible [19]. These structures provide valuable insights into the molecular basis of protein function, allowing an effective design of experiments such as site-directed mutagenesis, studies of disease-related mutations or the structure based design of specific inhibitors or drugs. However, the experimental determination of the 3D structure of a protein is a laborious task and thus, the number of structurally characterized proteins is low compared to the number of known protein sequences. Computational methods attempt to complement experimental techniques by predicting the structure of proteins. For example, comparative modeling predicts the three-dimensional structure of a given protein sequence based primarily on its alignment to one or more homologous proteins of known structure [20].

1.3.3 Gene Expression Profiling

DNA Microarrays are commonly used in expression analysis studies to monitor the expression levels of thousands of genes simultaneously [21]. These arrays consist of thousands of individual gene sequences bound to closely spaced regions on the surface of a glass microscope slide. There are two main types of DNA Microarrays that measure expression: two-channel microarrays and oligonucleotide microarrays. In a two-channel microarray, one single microarray can be used to visualize up-regulated and down-regulated genes in two different samples (e.g. patient and control), but absolute gene expression levels cannot be observed. Oligonucleotide microarrays give estimations of gene expression absolute levels and therefore the comparison of two conditions requires the use of two separate microarrays. Moreover, both types of microarrays can be used to estimate if two genes show similar expression behaviors (i.e. are coexpressed).

1.3.4 Protein chips

Protein microarrays are miniaturized and parallel assay systems that contain small amounts

of purified proteins in a high-density format [22]. There are two main classes of protein microarrays: analytical and functional protein microarrays [23]. On one hand, analytical protein arrays can be used to monitor protein expression levels or for biomarker identification. On the other hand, typical uses of functional protein microarrays include probing for various types of protein activities (e.g. catalytic activity) and to profile immune responses. Data from these experiments is of great use for proteomics studies. For example, protein chips have already been used to provide a global analysis of protein phosphorylation in yeast, which is a major regulatory mechanism that controls many basic cellular processes [24].

1.3.5 ChIp-on-Chip

ChIP-on-chip is a technique that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip) [25]. ChIP-on-chip is used to investigate interactions between proteins and DNA. These interactions mediate transcription, DNA replication, recombination and DNA repair, which are all fundamental to life. ChIP-on-chip experiments allow the determination of the entire spectrum of in vivo DNA binding sites for any given protein. However, this technique has still not been applied at proteome-scale levels, and currently, computational tools are being used to complement it by predicting DNA sites where proteins bind [26].

1.3.6 Fluorescent tagging

Fluorescent tagging is the process of attaching a fluorescent molecule (typically the green fluorescent protein (GFP) or a fluophore called fluorescein) to another molecule with the aim of aiding in detection of the molecule to which it has been attached. Fluorescent tagging applications include the analysis of protein expression patterns and determining the subcellular localization of proteins [27, 28]. Protein localization data obtained from these experiments are a valuable information resource helpful in elucidating eukaryotic protein function.

1.3.7 Methods for the detection of protein-protein interactions

Historically, biochemical approaches such as cross-linking, immunoprecipitation, and protein affinity chromatography have been used to verify interactions between suspected interaction partners [29]. However, these methods are not suited to analyze full proteomes in a reasonable time [29-31]. Recently, most approaches have been focused on large-scale screenings of protein-protein interactions, such as two-hybrid assays [5, 6, 32] and affinity

purifications or immunoprecipitation followed by mass spectrometry [33-36]. New approaches such as protein microarrays [22, 37] or luminescence-based mammalian interactome mapping (LUMIER) [38] have shown promise at detecting protein-protein interactions but haven't been still applied to proteome-scale mapping of interactions. Experimental techniques that detect large numbers of interactions by means of a single experiment are referred as "high-throughput methods".

In the **yeast two-hybrid assay** (Figure 1.3), a protein of interest (referred as 'bait') is typically fused to a DNA-binding domain (DBD). Other proteins (referred as 'preys'), which are fused to a transcription-activating domain (TAD), are screened for physical interactions with the bait protein using the activation of a transcription reporter construct as the detection method [39]. The interaction between the bait (fused to DBD) and the prey (fused to TAD) restores the function of the transcription factor, and activates reporter genes or selection markers.

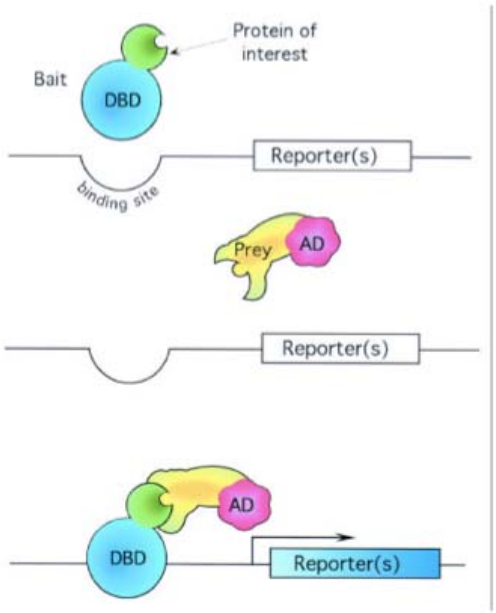


Figure 1.3 The yeast two-hybrid method. Image obtained from [40]

The **tandem affinity purification method** (TAP) requires fusioning a tag to the target protein (bait) of interest (Figure 1.4). The TAP tag often consists of calmodulin binding peptide (CBP), followed by tobacco etch virus protease (TEV protease) cleavage site and Protein A, which binds tightly to IgG. The target protein, fused to the TAP tag, is expressed in yeast, where it can form native complexes with other proteins (preys). After two steps of washing, the target protein complex is released from the IgG matrix, and the components of the complex are screened with mass spectrometry [33]. Moreover, binary protein

interactions can be inferred from the identified complexes using two types of interpretations [41]: (i) spoke: one interaction is defined between a bait protein and each protein it pulls down; (ii) matrix: interactions are defined between all pairs of proteins pulled down by a bait.

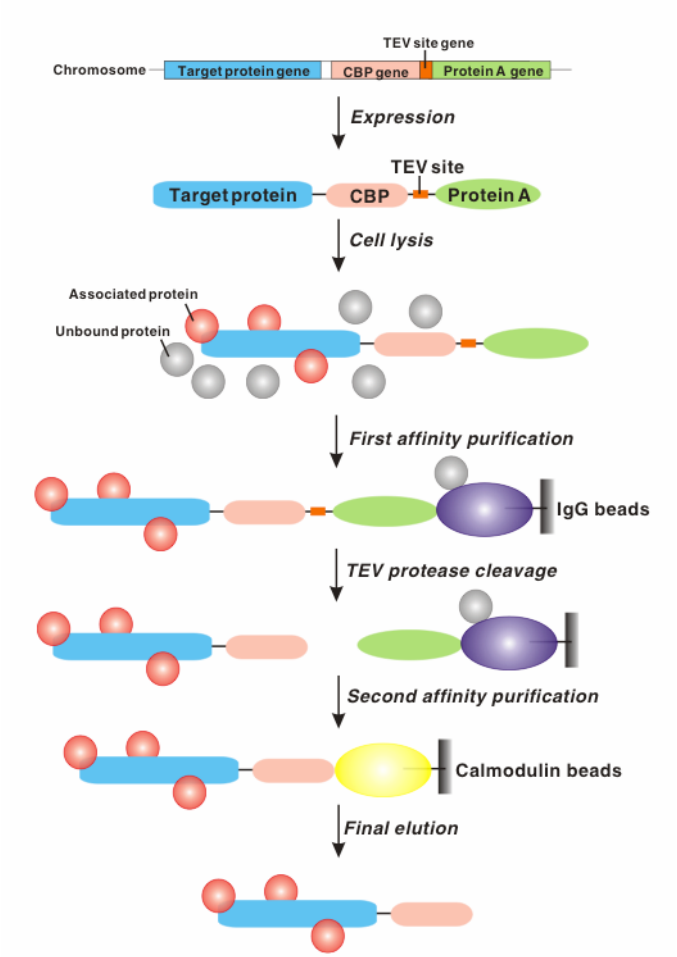


Figure 1.4 The Tandem Affinity Purification method. Image obtained from [42].

High-throughput methods (HT methods) for detecting protein-protein interactions have produced large amounts of data, but their reliability has been questioned [43-45]. Specifically, HT methods generate a lot of false positives (i.e. proteins that do not interact in vivo are reported to interact by the method), but also miss many interactions [41, 43]. Moreover, each technique is biased towards detecting and missing certain types of interactions and proteins [43]. For example, the TAP method is not suitable for screening transient protein interactions, unlike the yeast 2-hybrid method. However, TAP is a good method for testing permanent interactions and detects real complexes in physiological settings. Besides, while bait and prey have to be over-expressed in a yeast two hybrid system, the protein abundance in TAP is physiological. This might have an effect on the

detected interactions. For example, a number of false positives in yeast two hybrid are plausible interactions (i.e. the two proteins do interact when one is facing the other) that do not take place in vivo (e.g. they are never coexpressed in the cell) [41, 43]. It has also been shown that each technique produces a unique distribution of interactions with respect to functional categories of interacting proteins [41, 43].

Data of protein-protein interactions detected with experimental methods is being complemented (and filtered) using computational approaches [46-48]. Some common approaches to the prediction of protein-protein interactions include: (i) transferring interactions between protein orthologs of different species (referred as 'interologs') [49]; (ii) inferring interactions from correlated mutations [50]; (iii) predicting interactions using distant conservation of sequence patterns and structure relationships [51]; (iv) predicting interactions on the basis of co-occurrence of domains or sequence signatures [52]; (v) detecting gene fusion events [53]; and (vi) automatically mining scientific publications to detect interacting proteins [54]. Moreover, in addition to predicting physical interactions between proteins, other computational approaches attempt to create functional links between proteins using observations such as similar phylogenetic profiles, coexpression patterns or frequent gene neighborhood [55-57].

Computational approaches are also used to increase the confidence of interaction data [58]. A number of works have used the similarity of mRNA expression profiles to determine if an observed interaction has high, medium or low confidence [59, 60]. Several approaches have developed reliability measures based on the topology of the protein interaction network (see section 1.5.1), such as the "interaction generality" [61] and the IRAP* [62] methods. Other works apply an integrative approach in which they combine various features of interacting proteins, such as (i) functional similarity and high network clustering [63]; (ii) domain composition, Gene Ontology annotations [64] and sequence homology [65]; and (iii) statistical and topological descriptors, mRNA expression, genetic interactions and database annotations [66]. Finally, there are methods specifically designed to calculate the reliability of interactions detected by particular experimental methods, such as the socio-affinity index for TAP [35], which quantifies the tendency for proteins to identify each other when tagged (the spoke model) and to co-purify when other proteins are tagged (the matrix model).

The challenge in protein-protein interactions technologies resides not only in detecting a higher number of interactions and augmenting the reliability of the detection methods, but also in creating the tools that facilitate the correct storage, analysis and use of the interaction

data available. Moreover, due to the different nature of interactions and proteins detected by the methods, integrating data coming from multiple sources should be a fundamental goal of efforts dedicated to produce interactome maps.

1.4 Repositories of biological information

The scientific community needs to be able to retrieve the known biological information and use it for further computational and experimental scrutiny. Moreover, data from experimental studies is increasingly being made accessible for researchers willing to further analyze the results. One good example of a database that has become a de-facto standard for data storage is the Protein Data Bank [67], which holds all known protein 3D structures. However, the PDB is also a good example of a database that needs to adapt to the ever-increasing demands of the scientific community [68]. In this section, we briefly describe public repositories of biological information and experimental data.

1.4.1 Bio-sequences repositories

Various public repositories are dedicated to storing the knowledge available for different types of biomolecules, such as NCBI GenBank for genes [69] or UniProt for proteins [70]. Recently, we have observed a consolidation of these repositories, and cross-linking between the different records is increasingly becoming available, facilitating the access and use of data. However, the level of integration between the different repositories is still far from being optimal. For example, each database uses its own internal identifiers for biomolecules, and translating from one identifier to another is usually a daunting task [71].

1.4.2 Gene expression studies repositories

Microarray results are being stored in public repositories such the Gene Expression Omnibus [72] and the Array Express [73], and thousands of expression profiles are currently available. However, although there have been efforts to implement guidelines for expression data annotation and exchange [74, 75], a complete standardization of expression data has still not been achieved, which results in difficulties when trying to systematically analyze all data available.

1.4.3 Protein domains, families and functional sites

Proteins are known to be built from a limited set of molecular block types [76-79]. Several databases have been created that describe proteins in terms of structural domains, sequence domains and functional sites (Table 1).

Table 1. Repositories that classify proteins according to domains, families or functional sites.

Name	Available data	Description
SCOP [77]	75,930 domains, 3,004 families	SCOP comprehensively orders all proteins of known structure, according to their evolutionary and structural relationships. Protein domains in SCOP are hierarchically classified into families, superfamilies, folds and classes.
CATH [79]	86,151 domains	CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H)
ASTRAL [80]	54,745 domains	ASTRAL is a collection of databases and tools to aid in the analysis of protein structures, particularly through the use of their sequences. Partially derived from the SCOP database of protein structure domains, it includes sequences for each domain and other resources useful for studying these sequences and domain structures.
Pfam [81]	8,957 families	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families
INTERPRO [82]	13,828 entries, 3,905 domains, 9,614 families, 232 repeats, 34 active sites, 22 binding sites, 21 post-translational modification sites	InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. InterPro combines a number of databases that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures : PROSITE, Gene3D, PANTHER, PIRSF, Pfam, SMART, SUPERFAMILY and TIGRFAMs, PRINTS.
PROSITE [83]	1,327 patterns	PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.
PRINTS [84]	1,800 entries	PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.
SUPERFAMILY [85]	---	SUPERFAMILY provides structural (and hence implied functional) assignments to protein sequences at the superfamily level as defined by SCOP.
PRODOM [86]	736,449 families	ProDom is a comprehensive set of protein domain families automatically generated from the global comparison of all available protein sequences.

For example, the Structural Classification of Proteins (SCOP) [77] classifies protein

domains on hierarchical levels that embody the evolutionary and structural relationships (Figure 1.5). Proteins with a common evolutionary origin are grouped together in families. Proteins with low sequence identity whose structures and functional features suggest a common evolutionary origin are placed together in superfamilies. Finally, superfamilies and families are defined as having a common fold if their proteins have most of their secondary structures in the same arrangement and the same topological connections.

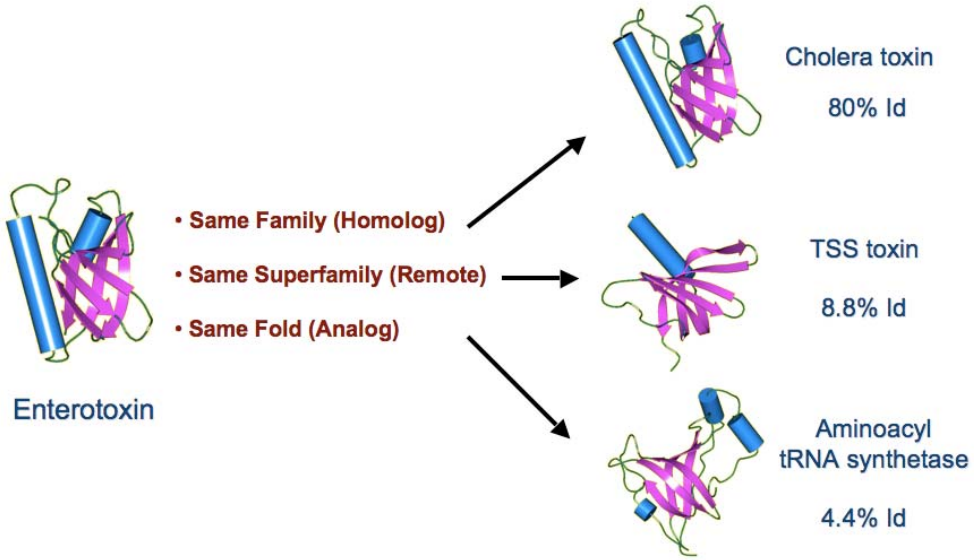


Figure 1.5. Example of domains within the same family, superfamily and fold.

1.4.4 Protein function repositories

Determining the functions of genes and proteins is a central problem in biology, fundamental to understanding the molecular and biochemical processes that sustain health or cause disease, to identifying and validating new drug targets and to developing reliable diagnostics [87]. To date, even for the most well-studied organisms such as yeast, about one-fourth of the proteome remain uncharacterized [88]. Functional classifications derive groups of genes and proteins on the basis of functional similarity in terms of enzyme reaction mechanisms, participation in biochemical pathways and functional roles [89]. These classifications provide a convenient framework for bioinformatics efforts geared towards protein function prediction.

The main classification schemes for protein function are the Enzyme Commission (EC) hierarchical classification [90] and the Gene Ontology (GO) [91]. In particular, the GO project provides structured, controlled vocabularies and classifications that cover several

domains of molecular and cellular biology. The controlled vocabularies (ontologies) describe proteins in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. These ontologies are structured as directed acyclic graphs, which are similar to hierarchies but differ in that a child -or more specialized- term can have many parents -or less specialized- terms. Moreover, a protein might be associated with or located in one or more cellular components; it might as well be active in one or more biological processes, during which it can perform one or more molecular functions.

Other repositories classify proteins (and other biomolecules) in terms of the biological pathways in which they intervene [92]. Two important databases of pathways are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [93] and REACTOME [94]. Originally created to store metabolic reactions, today they have been extended to contain as well other types of biological processes. Genmapp [95], another pathway tool, was specially designed to view and analyze microarray data on biological pathways.

Manually curated repositories of protein function are not always capable of keeping apace with the deluge of data coming from genomics and proteomics efforts [96, 97]. Thus, computational methods for annotating proteins are needed to complement experimentally validated data and manual curation. Computational approaches to the prediction of protein function [88, 98] include phylogenetic profiles [55], manual annotation from automatic literature search [91], transfer of function by sequence similarity [99], structure-based methods [100], integrative approaches that combine diverse functional genomics data [101] and network-based methods (see section 1.5.2). These approaches provide fundamental information about proteins for which the function has not already been experimentally determined.

1.4.5 Protein-protein interactions repositories

Results from protein-protein interactions screenings and predictions are being placed in public repositories of biological interactions, which enable a convenient access to the information available and facilitate further analyses on the data [92, 102, 103]. Table 2 describes the main repositories of protein interactions.

Table 2. Main repositories of protein-protein interactions (PPI).

Name	Number of Interactions	Type of Interactions
IntAct [104]	136,355	IntAct is the protein interaction database at EBI. No species restriction. Contains PPI from multiple types of sources (HT methods, direct submission, literature…).
DIP [105]	55,895	Database of Interacting Proteins (DIP) at UCLA. No species restriction. Contains PPI from multiple types of sources (HT methods, direct submission, literature…).
HPRD [106]	37,581	The Human Protein Reference Database (HPRD) contains human PPI extracted (and curated) from literature.
BIND [107]	>300,000	Biomolecular INteraction Network Database (BIND) at the University of Toronto, Canada. No species restriction. Archives biomolecular interaction, complex and pathway information. Contains PPI from multiple types of sources (HT methods, direct submission, literature…).
MIPS [108]	1,814	The MIPS Mammalian Protein-Protein Interaction Database is a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators. Contains mammalian PPI.
MINT [109]	102,571	The Molecular INTeraction (MINT) database contains PPI from multiple types of sources (HT methods, direct submission, literature…).
BioGrid [110]	167,752	The General Repository for Interaction Datasets (BioGrid) is a database of genetic and physical interactions. Contains PPI for 13 organisms from multiple types of sources (HT methods, direct submission, literature…).
CYGD [111]		CYGD is the PPI section of the Comprehensive Yeast Genome Database. Contains manually curated comprehensive S. cerevisiae PPI.
STRING [112]	730,000 proteins	The Search Tool for the Retrieval of Interacting Proteins (STRING) covers 1.5 million proteins for 373 species. Features AJAX-based web-navigation, inclusion of BioGRID, and detailed protein domain annotation. It is useful for comparative genomics, phylogenetics and network studies. Contains PPI from multiple types of sources (HT methods, direct submission, literature, predictions…).
HomoMINT [113]	--	HomoMINT contains inferred human PPI from orthology to model organisms
HPID [114]	--	The Human Protein Interaction Database (HPID) contains predictions of human interaction derived from model organisms.

Most of these databases keep interactions as binary relationships between proteins. In

addition, binary interactions are usually annotated with: (i) detection method employed; (ii) Pubmed identifier of the article where the interaction was described; and (iii) several protein identifiers for the two proteins involved in the interaction. Some other databases also include information such as complex membership, role (bait/prey) in the detection method, or a confidence score for the interaction. Most databases contain interactions obtained by direct submission from experimentalists and by mining literature and other data sources; in some cases the data is verified using automated algorithms or manual curation [103].

Additionaly, recent works have focused on creating repositories of interactions between protein interfaces. In these repositories, interactions are described at resolutions lower than the whole protein, usually as domain-domain interactions. PIBASE [115] is a comprehensive relational database of structurally defined domain-domain interfaces, based on domains definitions from SCOP [77] and CATH [116]. SCOPPI [117] classifies and annotates domain interactions derived from all known protein structures, applying SCOP domain definitions [77]. Another resource, iPfam [118] investigates protein interactions in the Protein Data Bank (PDB) [67] at the level of Pfam domains [81] and aminoacid residues. Moreover, numerous computational methods have been used to complement interface interactions extracted from the PDB by predicting domain-domain interactions from known protein interactions and classifications of protein domains. For example, Sprinzak and Margalit characterized proteins using InterPro structural domains and then used experimentally determined protein-protein interactions to identify structural domain pairings that correlate with protein binding [52]. Other approaches characterized proteins using evolutionarily conserved domains defined in Pfam and predicted domain-domain interactions by applying a maximum likelihood method [119] or using a probabilistic confidence scoring scheme to combine multiple data sources [120].

In conclusion, protein interaction data can be found in repositories of proteins and domain interactions, which might contain results from multiple experimental studies or computational predictions. In spite of the interaction data diversity, recent standardization efforts (see section 1.6.2) have increased the overlap between the different databases. However, there is no one definitive database for protein-protein interactions, and integration efforts that unify all available data are needed.

1.5 Protein interaction networks

Many complex systems can be represented and analyzed as networks [121]. In particular, the network analysis approach is fundamental for successful quantitative modeling of biological systems [122, 123]. Network-based approaches have been used to analyze biological systems such as gene regulatory networks [124], signal transduction networks [125], metabolic networks [126], gene co-expression networks [127], protein interaction networks [128], phenome-genome networks [129] and phenome-interactome networks [130].

Protein-protein interactions is one type of biological data that can be represented and analyzed as a network [8]. In a protein interaction network, the nodes are proteins and the edges represent physical interactions between proteins. Formally, a protein-protein interaction network is defined as a set of proteins $P=[p_1,...,p_n]$ with interactions $I=[i_{11}, \ldots, i_{nm}]$ between them, where i_{jk} describes an interaction between proteins j and k. In such a network, a set of proteins linked to protein p_j (ie, physically interacting with p_j) is named "partners of p_j". The distance between two proteins of the network is defined as the minimum number of edges that one has to follow in order to connect the two proteins. The protein-protein interaction network for a given protein can be built at different depths, which represents the number of interacting steps that can be taken from the source protein to the outermost protein of the network. Consequently, the protein interaction network for a given species (i.e. interactome network) is the joint network of all proteins within that species. For example, building a network at depth 2 for a particular protein p_j implies adding to the network the parters of p_j as well as the partners of the partners of p_j. In a protein interaction network, we refer to proteins with high connectivity (i.e. with many interaction partners) as 'hubs'. Figure 1.6 shows the recently published protein interaction network for *Drosophila Melanogaster* [3]. As illustrated in Figure 1.6, sections 1.5.1 and 1.5.2, representing protein interactions in a network has fundamental advantages over the traditional approach of storing interaction data in the form of simple lists.

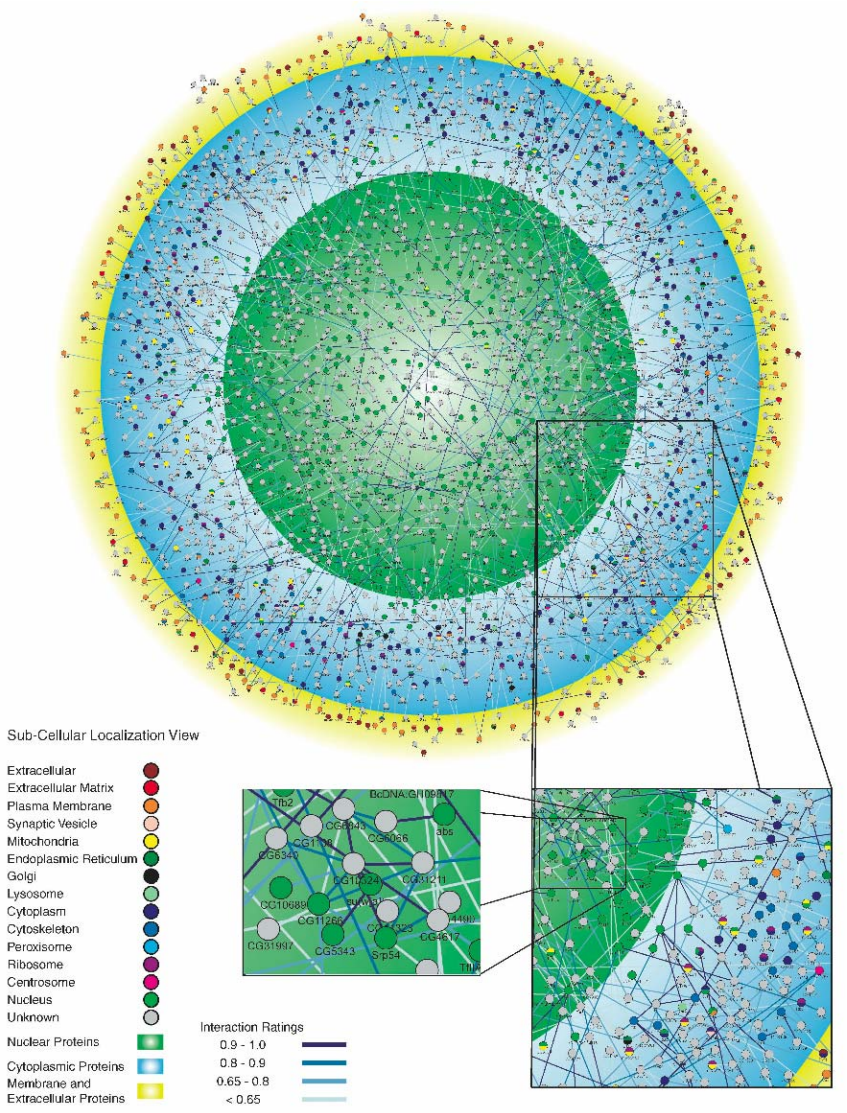


Figure 1.6 Protein-protein interaction network of the fruit fly from a cellular localization perspective. Image from [131]

1.5.1 Protein interaction networks properties

There are indications that protein interaction networks are governed by certain universal laws [8]. For example, interactome networks appear to have a 'scale-free' or power law degree distribution: most proteins interact with few partners, whereas a few proteins, named 'hubs', interact with many partners [8, 132, 133]. Nevertheless, the scale-free aspect of networks will need to be revisited when more comprehensive interaction networks are compiled, since a number of studies have recently questioned its existence [134, 135]. Besides, a link between the potential scale-free topology of interactome networks and genetic robustness seems to exist, because it has been shown that hub proteins are more essential for cellular viability than non-hubs [128]. However, recent studies have suggested

that the essentiality of protein hubs is better explained by their number of binding interfaces [136], the type of interactions in which hubs are involved [137] or the number of shortest paths going through them [138].

Other properties observed in interaction networks include the small-world effect (i.e. any two nodes can be connected with a path of a few edges only) and a highly modular architecture (i.e. interacting proteins tend to be in complexes or act in the same functional process) [8, 139-141].

1.5.2 Protein interaction networks applications

Protein interaction networks have been shown to be a useful representation of interaction data [142]. Moreover, networks are useful for integrating data from multiple sources (see section1.6.2). Furthermore, protein interaction networks can be applied to objectives such as (i) the prediction of protein function [88]; (ii) the estimation of interactions reliability (see section 1.3.7) [66]; (iii) the identification of domain-domain interactions (see section 1.4.5) [143]; (iii) the prediction of protein interactions (see section 1.4.5) [144]; (iv) the detection of proteins involved in disease pathways (see section 1.8) [145]; (v) the delineation of frequent interaction network motifs [146]; and (vi) the comparison between model organisms and humans [147].

In particular, protein interaction networks have been extensively used to predict the function of proteins [88, 148]. Network-based approaches for elucidating protein function can be classified in direct methods, which propagate functional information through the network, and module-assisted networks, which infer functional modules within the network and use those for the annotation task [88]. The common principle underlying all **direct methods** for functional annotation is that proteins that lie closer to one another in the PPI network are more likely to have similar function. Examples of direct methods are neighborhood counting [149], graph theoretic approaches that take into account the full topology of the network [150] and Markov random field probabilistic approaches [151]. **Module-assisted methods** attempt to identify coherent groups of genes (in terms of network topology or data from experiments) and then assign functions to all the genes in each group. One interesting example of module-assisted method is the work of Samanta and Liang [152], which transfers function between proteins that have a significant number of common interaction partners.

Finally, two other important applications of protein interaction networks are the prediction of protein-protein interactions and the identification of frequent network motifs.

Network approaches to the prediction of protein interactions are usually based on inferred domain-domain interactions (see section 1.4.5) or topological properties of protein networks, such as conserved interaction patterns [153]. Methods based on interaction patterns rely of the observation that certain network motifs such as a triad of interactions occur in protein interaction networks at a significantly higher frequency than that expected from an artificially generated network with similar mathematical properties [132]. These conserved network motifs can then be used to generate plausible predictions of protein-protein interactions.

1.6 Biological data integration

Over the past two decades, databases of biological knowledge have become essential resources that are daily used by biologists around the world. One fundamental problem in using these databases is that, in order to answer any biological question, it is necessary to traverse several repositories of knowledge, because information is spread across multiple databases, websites, text files, and private repositories of data [154-156]. Moreover, biological information is stored in multiple (disparate) formats, increasing the difficulty when trying to uniformly use all available information. Furthermore, the knowledge stored relies on different nomenclatures: the same biological object (e.g. a protein) might be identified with a different name in each repository that contains information about it.

Several consequences arise from the above mentioned issues: (i) biologists won't find the information they are searching unless it appears in their favorite websites; (ii) since having access to all available information on a particular subject (e.g. protein-protein interactions) is not straightforward, only one source of data will be used for testing a method or checking the validity of a hypothesis; and (iii) analyses based on multiple sources of data are subjected to errors, due to different nomenclature systems and lack of one-to-one translations. Thus, solving these difficulties through effective data integration is a key element of conducting biological research [154, 157].

Ideally, one single universal database would contain all types of biological data. However, this is not feasible in practice because (i) creating a single data model for all types of information is difficult (if not impossible); (ii) different groups hold the diverse expertises; and (iii) keeping up to date repositories is easier when the data to be maintained is limited. A more practical solution is to maintain the scientific and political independence

of the databases, but make the information that they contain easy to integrate by means of cross-database queries based on exchange standards. However, this is not trivial, for both technical and human-dependent issues [154, 158].

There are three main ways in which groups have tried to integrate biological databases: link integration, view integration and data warehouses [154]. **Link integration** is the simplest (and most widely used) approach to having access to all available data: researches begin their query with one data source, and then follow hypertext links to related information in other data sources. **View integration** leaves the information in its source databases, but builds an environment around the databases that makes them all seem to be part of one large system [159]. **Data warehousing** (Figure 1.7) consists in integrating all the data into a single database, which can then be used as the 'one-stop shop' for answering any of the questions that the source databases can handle, as well as those that require integrated knowledge that the individual sources do not have [160].

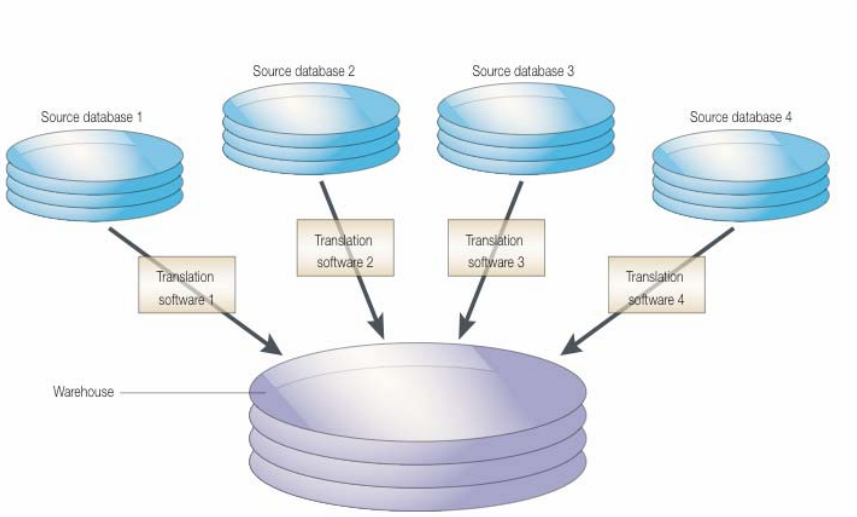


Figure 1.7 The data warehousing approach to data integration. Image obtained from [154]

This section addresses the challenge of integrating biological data, focusing on aspects related to the study of protein-protein interactions. Specifically, we discuss the nomenclature issues in biology, the integration of protein-protein interactions and the combination of protein-protein interaction data with other types of biological data.

1.6.1 Genes and proteins nomenclature

Adopting naming standards and translating between different biological entities identifiers is the first step towards biological data integration [71, 161]. As described in section 1.2, genes are transcribed into mRNA, which is translated to proteins. However, numerous mechanisms affect the process of making proteins, such as alternative splicing (i.e. process that occurs in eukaryotes in which the splicing process of a pre-mRNA can lead to different ripe mRNA molecules and therefore to different proteins) or post-translational modifications (e.g. phosphorylation, glycosylation, acetylation). Therefore, a gene is a recipe for one or more proteins. However, most biologists use gene names when referring to proteins. Moreover, biologists use a variety of names for genes and proteins, based on factors such as their research specialization, historical reasons or taste for fanciful names [162]. This brings great difficulties to the process of searching for information for a given gene, and most pronouncedly, to efforts dedicated to integrating information coming from multiple sources. Some examples of gene naming issues are: (i) multiple names can be used to refer to the same gene (e.g. CAPZA3, CAPAA3 and GSG3 all refer to F-actin capping protein subunit alpha-3); (ii) the same gene name can be written in different ways (e.g. spaces become hyphens); (iii) the same gene name can refer to several unrelated proteins (e.g. searching in UniProt for gene name 'pap' retrieves 57 different proteins, many of them completely unrelated); and (iv) gene names can be species specific (e.g. PMS1 is the yeast ortholog of human PMS2; the yeast ortholog of human PMS1 is PMS2).

Several efforts have been made to rationalize the naming of genes and proteins. In particular, standards that guarantee the use of unique names are being adopted in most commonly used organisms [163-166]. Moreover, there are as well databases that have developed unique and tractable identifiers such as Uniprot Entry Names [167], Entrez GeneIDs [69] and Ensembl [168]. However, attempts to impose standard names are meeting stiff resistance, and approaches that use unique identifiers are not practical for every day use. In order to address the nomenclature problem, a number of bioinformatics approaches have been developed to automatically translate between different identifiers (see section 1.7.1).

1.6.2 Protein-protein interactions integration

There are two aspects to protein-protein interactions integration. The first aspect is collecting all sources of interaction data and creating a unified interaction network. The second aspect is analyzing protein-protein interaction data in combination with other types of biological data.

Integrating protein-protein interactions from multiple sources

Using protein-protein interactions for computational biology analyses has been historically plagued with problems unrelated to biological questions. For example, disparate formats and ambiguous protein identifiers impeded the compilation of a comprehensive list of known interactions, making difficult to prove (or disprove) research hypothesis. Traditionally, researchers kept and exchanged protein-protein interaction data in simple text files. However, this approach was not suited for the deluge of data produced by high throughput methods. Consequently, public repositories of protein-protein interactions were created, either by direct submissions of interactions or manual curation of interactions from the literature. Recently, one standard has emerged that attempts to uniform the way interactions are formatted and codified [169]. The proposed standard, the protein standards initiative for molecular interactions (PSI-MI), is a data exchange format, not a proposed database structure. This standard is being promoted by the International Molecular Exchange Consortium (IMEx)[170], a group of major public interaction data providers sharing curation efforts and exchanging completed records on molecular interaction data. The two main contributions of the PSI-MI are the definition of a XML specification for exchanging molecular interaction data and the creation of controlled vocabularies (e.g. fixed identifiers for types of detection methods).

While the PSI-MI has standardize to a certain degree the way interactions can be exchanged between different databases, using all available interactions in an integrated manner is still out of reach for most computational biologists, not to say for biologists in experimental labs. The main causes for this are: (i) protein nomenclature issues are there to stay (see section 1.6.1); (ii) databases do not follow the PSI-MI in a uniform manner (including those that promote the standard); (iii) a unified network of all known interactions has not been compiled yet into a single repository. In consequence, researchers willing to work with all available interactions need to recur to third party integration tools (see section 1.7.2).

The integration of protein-protein interactions from multiple sources, in addition to providing a greater coverage of the interactome space [171, 172], can be helpful for obtaining high confidence networks (see section 1.3.7). For example, interactions detected in multiple experiments are more reliable than those detected in just one experiment; the same is true for interactions detected with different experimental techniques [43, 171].

Combining protein-protein interaction data with other types of biological data

There are two main objectives when combining protein-protein interaction data with other types of biological data. The first objective is to increase the confidence of protein interaction data by determining the reliability of an interaction on the basis of other biological evidence. The second objective is to provide additional biological insights by extracting knowledge from the combination of biological data types, knowledge that each individual data source didn't contain by itself.

Methods for detecting protein-protein interactions have inherently high false-positive and false-negative rates (see section 1.3.7 and [43]). One way of **reducing the number of false positives** is to use other types of biological data to judge whether interactions are likely to occur in the cell. For example, localization data (see section 1.3.6) can be helpful for filtering false positives from protein-protein interactions detected by high-throughput methods: if two proteins are never co-localized, they are not likely to interact in vivo. Moreover, various protein/gene pair characteristics, such as shared phenotypes, correlated expression or shared GO terms, have been employed to assign higher confidence to protein-protein interaction data [5, 45].

Integrating protein-protein interactions with other types of data can provide valuable **biological insights** not captured by protein-protein interaction data alone [173]. For example, gene expression studies (see section 1.3.3) can be useful for (i) detecting modules of proteins that are coexpressed [174]; (ii) identifying network proteins that are differentially expressed in a given disease [145]; and (iii) viewing the interaction network from a dynamic perspective [175]. Another example of data that can be integrated with protein-protein interactions are results from ChIP-on-chip experiments (see section 1.3.5). In particular, combining information of protein-DNA binding sites with protein interactions could be fundamental to throwing light on transcription networks [176]. Finally, one example of integration of multiple types of biological data is the work of Tanay *et al,* in which they described a framework that allows the integration of protein interaction data with gene expression, phenotypic sensitivity and transcription factor binding [177].

1.7 Software

In this section we review some software packages related to the integration, analysis and visualization of protein-protein interaction data, including tools for translating between different types of protein identifiers.

1.7.1 Software for translating gene and protein identifiers

In order to address the nomenclature problem (see section 1.6.1), a number of bioinformatics approaches have been developed to automatically translate between different identifiers. Most translation tools rely on a data warehouse approach where information from multiple sources is stored. We describe here a summary of the main translation tools:

- *HomGL*: HomGL [178] is a php/mysql driven web-tool to compare and transfer genelists between different types of accession numbers and organisms (up to now: rat, human, mouse) utilizing the Unigene Clusters, the HomoloGene database, and the LocusLink database.
- *Onto-translate*: Onto-Translate [179] allows arbitrary mappings between 28 types of IDs for 53 organisms.
- *Gpsdb*: The Gene and Protein Synonym DataBase (GPSDB) [180] contains 552,469 gene and protein clusters, with an average of 24.2 names per cluster.
- *GeneSeer*: GeneSeer [162] allows access to gene information through common names and can map sequences to names. It also allows identification of homologs and paralogs for a given gene. GeneSeer works by collecting data from a variety of sources and bulding a name-translation database.
- *MatchMiner*: MatchMiner [181] is a freely available program package for batch navigation among gene and gene product identifier types commonly encountered in microarray studies and other forms of 'omic' research.
- *BioThesaurus*: BioThesaurus [182] is a web-based system designed to map a comprehensive collection of protein and gene names to protein entries in the UniProt Knowledgebase.
- *Connect the Dots*: Connect the Dots [183] is a general data integration tool that is focused on translating identifiers between public biological databases.

1.7.2 Software for protein interaction data integration, analysis and visualization

In the last five years, numerous software packages have been developed to integrate, analyze and visualize protein-protein interaction data. In this section, we describe the main actors in the molecular interaction software landscape.

- *Cytoscape*: Cytoscape [184] is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data. Figure 1.8 illustrates the user interface of Cytoscape.
- *VisANT*: VisANT [185] is an application for integrating biomolecular interaction data into a cohesive, graphical interface
- *Osprey*: Osprey [186] is a software framework for visualization of complex interaction networks
- *cPath*: cPath [187] is an open source database and web application for collecting, storing, and querying biological pathway data
- *ProViz*: ProViz [188] is a tool for the visualization of protein-protein interaction networks
- *PimWalker*: PimWalker [189] is a free and interactive tool for visualising protein interaction networks. PIMWalker handles the unified molecular interaction (MI) format defined by members of the Proteomics Standards Initiative (the PSI MI format)
- *tYNA*: tYNA [190] is a web system for managing, comparing and mining multiple networks, both directed and undirected
- *iVici*: iVici [191] is a Java application for analysis of cellular networks represented as addressable symmetric or asymmetric two-dimensional matrices.
- *InterViewer*: InterViewer [192] is a extremely fast layout algorithm for visualizing large-scale protein interaction networks.
- *Integrator*: Integrator [193] is a collection of interactive, graphical search tools for exploring protein-protein interaction networks.
- *AVID*: Avid [194] is a computational method that uses a multi-stage learning framework to integrate experimental results with sequence information, generating networks reflecting functional similarities among proteins.

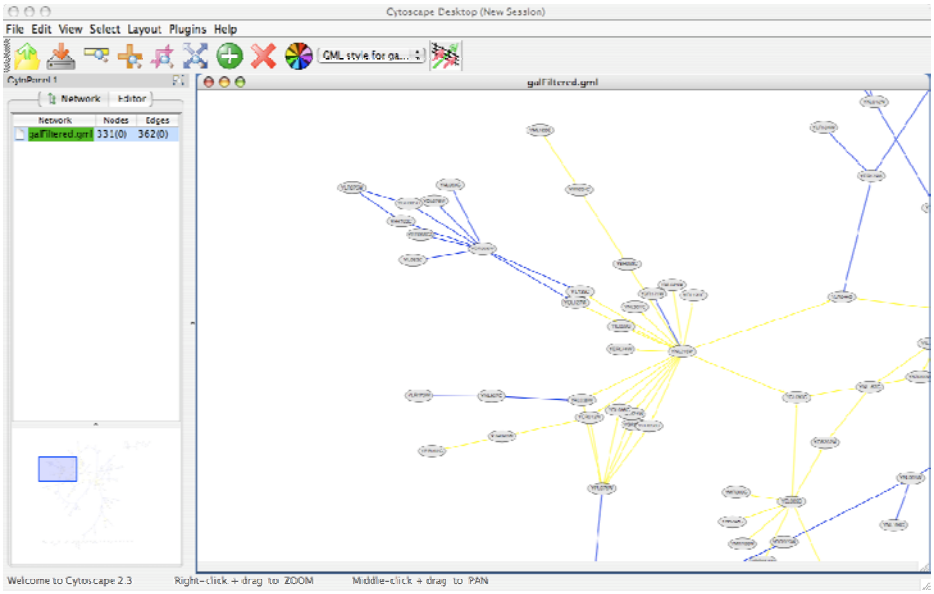


Figure 1.8 Screen capture of the user interface provided by Cytoscape [184].

Moreover, most databases listed in section 1.4.5 provide a web-based user interface that can be use to navigate their own protein interaction network. Furthermore, there are some public repositories that attempt to collect their data from as many sources as possible, including other databases. Two examples of this integrative approach are STRING [112] and UniHI [195].

1.8 The role of bioinformatics in cancer research

Bioinformatics is increasingly becoming a discipline with real life medical applications [196] [197]. Biomedical areas that have benefited from advances in bioinformatics research include drug design [198, 199], disease diagnosis [200, 201], identification of biomarkers for patient monitoring [202] and drug discovery [203, 204]. In particular, computational approaches are being used to speed up the identification of effective new drugs [203]. There

are three major issues associated with drug discovery: (i) identifying relevant drug targets; (ii) identifying a drug that appropriately disturbs the target; and (iii) assessing the possible side effects and pharmaceutical properties of the drug. Computational approaches are well suited for helping with issues (i) and (iii) [203], applying techniques such as virtual screening [204] and probabilistic modeling of human systems biology [205] .

One area where bioinformatics is helping to narrow the gap between basic research and clinical application is cancer diagnosis and treatment [206-208]. Cancer is a complex disease, involving multiple and specific changes at the DNA level that can be inherited or induced by environmental factors [209]. Identifying cancer-specific molecular changes and discovering their use for increasing therapeutic specificity will lead to higher success rates and fewer side effects coming from aggressive cancer treatments [206, 210]. With this objective, numerous genomics and proteomics projects are dedicated to the study of genes and proteins and their involvement in cancer [211]. Data from these projects is increasingly being used by bioinformatics approaches to improve cancer diagnosis, prognosis, prevention and therapy [206, 207].

In particular, there are two main types of data that are being widely use by the bioinformatics research community: microarray data and protein-protein interaction data. For example, computational methods have been useful for simultaneously analyzing multiple expression studies [212], classifying tumor types on the basis of their gene expression profiles [213], or identifying cancer genes by their topological features in the protein interaction network [214]. Recently, several approaches have developed algorithms that perform integrated analyses based on heterogeneous sources of information such as gene expression studies and protein-protein interactions [215]. In the work of Rhodes *et al*. [145], a human interactome was applied to genome-wide gene expression data in cancer for identifying a potential tumor suppressor gene in the integrin signaling pathway, and then demonstrated the utility of protein-protein interaction data for identifying interaction subnetworks activated in cancer.

Besides, bioinformatics is not only playing an increasingly important role by developing computational methods and tools, but also for the continuing work of creating standards and repositories [207]. Public data sources with cancer-related information include Oncomine, a resource for examining gene expression in cancer [216]; the Cancer Gene Census, a catalogue of genes for which mutations have been causally implicated in cancer[217]; and the cancer ‘module map’, which describes expression profiles in different tumors in terms of the behavior of modules, sets of genes that act in concert to perform a specific function

[213]. Moreover, most repositories described in section 1.4.5 contain information that is potentially useful for the study of cancer.

1.9 Document organization

This PhD thesis is presented as a compendium of published (or submitted) research articles. The research articles are accompanied by an introduction (this chapter) that puts our work into context, and a discussion (chapter VI) that summarizes the achievements of the presented work. Specifically, this thesis is organized in the following chapters:

Chapter I (this chapter) introduces the basic biological concepts needed for reading this thesis and provides an overview of experimental methods, repositories of biological data, protein interaction networks, biological data integration and the application of bioinformatics to cancer research.

Chapter II introduces PIANA by means of two articles: one focused on the application (published on Bioinformatics as an Application Note) and another one dedicated to the data integration approach of PIANA and a description of the interaction data currently available (submitted to BMC Bioinformatics as a research article).

Chapter III describes a method implemented in PIANA that attacks the problem of not being able to know the interacting interfaces of proteins by means of high-throughput methods. In the article included in Chapter III (in revision in PLoS Computational Biology), we used binary protein interactions to identify proteins that interact through common interacting motifs. These interacting motifs were then used to show that the essentiality and evolutionary rate of hub proteins are related to their number of distinct interacting motifs, rather than to the number of interaction partners.

Chapter IV describes the use of PIANA from the biomedical perspective of identifying cancer gene candidates by the integration of multiple sources of data. The article included in Chapter IV (submitted to BMC Bioinformatics as a research article) predicts cancer gene candidates by integrating a list of known cancer genes, protein-protein interaction data, differential expression studies and structural and functional properties.

Chapter V is a summary of other works where the author of the thesis has been involved, such as publications from an on-going collaboration with wet lab experimentalists working in cancer and a research article in which we combine sequence information with protein-protein interactions to detect remote homologs (published in PNAS).

Finally, **Chapter VI** suggests future directions of research and summarizes the work presented in this PhD thesis.

OBJECTIVES

The objectives of this PhD thesis can be summarized as follows:

1. Designing and implementing a software platform for integrating, managing and analyzing protein-protein interactions.
2. Using protein-protein interaction data to provide biological insights about the mechanisms behind protein interactions. Specifically, we were interested in describing protein-protein interactions in terms of the specific interfaces that are in contact during an interaction.
3. Using protein-protein interaction data from a biomedical perspective. In particular, we were interested in demonstrating that existing data on protein-protein interactions can be of use for predicting proteins involved in disease.

The first objective was addressed by designing, implementing and publicly releasing a software framework called PIANA (Protein Interactions and Network Analysis) (see chapter II). The second objective was pursued by developing a method within PIANA that infers interacting motifs from binary protein interactions (see chapter III). The third objective was tackled with the application of PIANA to the prediction of cancer genes by integrating protein interaction networks and differential expression data (see chapter IV).

LIST OF PUBLICATIONS

Espadaler J[*], Aragues R[*], Eswar N, Marti-Renom MA, Querol E, Aviles X, Sali A, Oliva B. **Detecting remotely related proteins by their interactions and sequence similarity.** Proc Natl Acad Sci U S A. 2005 May 17;102(20):7151-6

[*] *Both authors contributed equally to this work*

Aragues R, Jaeggi D, Oliva B. **PIANA: protein interactions and network analysis.** Bioinformatics. 2006 Apr 15;22(8):1015-7

Mendez O, Martin B, Sanz R, Aragues R, Moreno V, Oliva B, Stresing V, Sierra A. **Underexpression of transcriptional regulators is common in metastatic breast cancer cells overexpressing Bcl-xL.** Carcinogenesis. 2006 Jun;27(6):1169-79

Espana L, Martin B, Aragues R, Chiva C, Oliva B, Andreu D, Sierra A. **Bcl-x(L)-mediated changes in metabolic pathways of breast cancer cells: from survival in the blood stream to organ-specific metastasis.** Am J Pathol. 2005 Oct;167(4):1125-37

Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. **Protein Hubs Characterization by Inferring Interacting Motifs from Protein Interactions.** PLoS Computational Biology (in revision)

Martín B, Aragues R, Sanz R, Oliva B, Boluda S, Martínez A, Sierra A. **Biological pathways contributing to organ-specific phenotype of brain metastatic cells** *Submitted to Oncogene*

Aragues R, Sander C, Oliva B. **An integrative approach to predicting cancer genes.** *Submitted to BMC Bioinformatics*

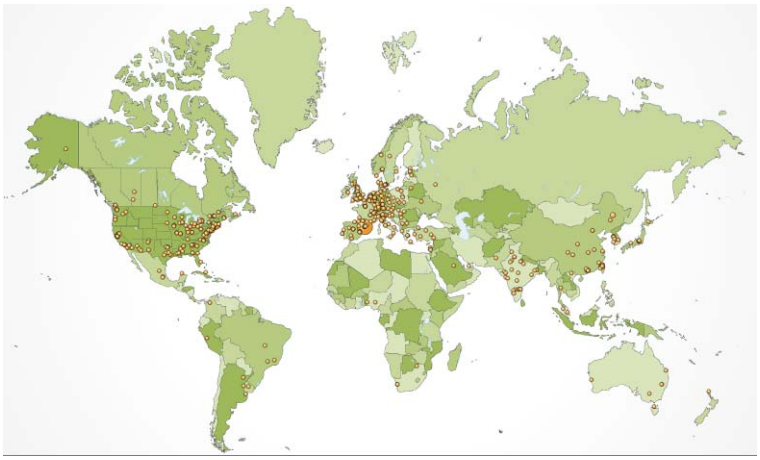
Aragues R[*], García-García J[*], Oliva B. **Assessment of protein-protein interaction data in the public domain by integration of multiple sources** *Submitted to BMC Bioinformatics*

[*] *Both authors contributed equally to this work*

CHAPTER II

PIANA
(Protein Interactions And Network Analysis)

As presented in the introduction to this PhD thesis, protein-protein interaction data is spread across multiple repositories, websites and literature. This hinders the advance of research in scientific areas where comprehensive access to protein-protein interaction data is a must. In this chapter, we present by means of two articles a software framework called PIANA designed to facilitate the analysis of protein interactions retrieved from multiple sources.



Access statistics to the PIANA website mapped on a world map.
As observed, the PIANA access map is a good surrogate for countries with a relevant scientific community.

Articles included in this chapter:

Aragues R, Jaeggi D, Oliva B. **PIANA: protein interactions and network analysis.** Bioinformatics. 2006 Apr 15;22(8):1015-7

Aragues R*, García-García J*, Oliva B. **Assessment of protein-protein interaction data in the public domain by integration of multiple sources** *Submitted to BMC Bioinformatics*

* *Both authors contributed equally to this work*

PIANA: Protein Interactions and Network Analysis

Ramon Aragues, Daniel Jaeggi and Baldo Oliva*

Structural Bioinformatics Group (GRIB –IMIM). Departament de Ciències Experimentals i de la Salut. Universitat Pompeu Fabra. Barcelona, Catalonia (Spain).

ABSTRACT

Summary: We present a software framework and tool called PIANA that facilitates working with protein interaction networks by 1) integrating data from multiple sources, 2) providing a library that handles graph-related tasks and 3) automating the analysis of protein-protein interaction networks. PIANA can also be used as a stand-alone application to create protein interaction networks and perform tasks such as predicting protein interactions and helping to identify spots in a 2D electrophoresis gel.

Availability: PIANA is under the GNU GPL. Source code, database and detailed documentation may be freely downloaded from http://sbi.imim.es/piana

Contact: ramon.aragues@upf.edu and boliva@imim.es

1 INTRODUCTION

The analysis of protein interaction networks is fundamental to the understanding of cellular processes (Salwinski and Eisenberg, 2003; Yook, et al., 2004). Furthermore, protein interaction networks are being used in tasks such as assignment of function to uncharacterized proteins (Huynen, et al., 2003) and searching for remote similarities between proteins (Espadaler, et al., 2005). Some tools developed to visualize and analyze protein-protein interaction networks are Cytoscape (Shannon, et al., 2003), Osprey (Breitkreutz, et al., 2003), VisANT (Hu, et al., 2005), and ProViz (Iragne, et al., 2005). Most of these tools are focused on visualizing the networks, while a few of them have analytic capabilities.

PIANA (Protein Interactions And Network Analysis) is a software framework that integrates data from multiple sources into a single repository, creates interaction networks, predicts novel interactions and performs automatic analyses. PIANA is different to most

*To whom correspondence should be addressed.
Grup de Bioinformàtica Estructural (GRIB-IMIM). Universitat Pompeu Fabra. C/ Doctor Aiguader, 83. Barcelona, 08003. Catalonia, Spain.

other tools in that 1) it is also a framework on which developers can base their applications, 2) it integrates most protein and interaction databases into a single repository, 3) it performs analyses not provided by other tools.

2 PIANA ARCHITECTURE

PIANA has been implemented as a collection of python modules that can be used separately as libraries or as a stand-alone application through a user interface.

The Database Module

The Database Module consists of a MySQL database and a library used as an interface to the database. A limited version of a PIANA MySQL database containing interactions from DIP (Salwinski, et al., 2004) and interactions predicted from sequence/structure distant patterns (Espadaler, et al., 2005) can be downloaded from our website.

The Parsing Module

PIANA includes parsers for the main protein databases (UniProt (Bairoch, et al., 2005), NCBI GenBank (Benson, et al., 2005)) and for protein interaction repositories such as DIP, STRING (von Mering, et al., 2003), MIPS (Pagel, et al., 2005), BIND (Alfarano, et al., 2005) and HPRD (Peri, et al., 2003). PIANA can also parse flat text files and interaction data that follows the HUPO PSI MI standard (Hermjakob, et al., 2004). Moreover, PIANA provides parsers for databases such as COG (Tatusov, et al., 2003), GO (Ashburner, et al., 2000) and SCOP (Murzin, et al., 1995). These databases contain information that PIANA uses when performing the analyses.

The Network Module

PIANA implements classes and methods for working with networks. Moreover, PIANA has methods specifically designed for biological networks such as clustering proteins by their molecular function and visualizing the networks in formats appropriate for biological analysis.

3 PIANA CAPABILITIES

Data integration

PIANA accepts as input most types of protein database identifiers and contains cross-references between them. Therefore, interactions from different sources can be integrated into a single network. Currently, the type of input and output protein database identifiers accepted by PIANA are: UniProt entry names and accession numbers, gene names, NCBI GenBank gi, EMBL, PDB, PIR and the protein sequence.

Creation of protein-protein interaction networks

Usually, a list of proteins of interest is given as input. PIANA searches in its database for interactions where these proteins are involved and adds edges (*ie* interactions) and nodes (*ie* protein interaction partners) to the network until a given depth is reached, where depth is defined as the number of interacting steps taken from the original proteins. Internally, a protein interaction network is represented as a set of nodes (proteins) connected by edges (interactions). The networks can be visualized in different formats, mainly tables that describe in detail each interaction and DOT files, which can be used to produce network images. PIANA also has the possibility of applying output filters such as highlighting proteins that perform specific functions or identifying proteins in the network whose genes have been found over/under expressed in a microarray experiment.

Predicting novel interactions

PIANA transfers interactions between proteins that share a given property. For example, PIANA predicts interactions using "interologs" (Yu, et al., 2004) by means of COG codes. In a similar way, SCOP codes can be used to transfer interactions between proteins that share a domain family.

Finding "interaction distance" between proteins

PIANA can obtain lists of proteins that are at a certain interaction distance (*ie* minimum number of edges separating two proteins) from another protein, which can be useful for tasks such as searching for remote similarities between proteins (Espadaler, et al., 2005).

Matching spots from electrophoresis experiments

PIANA can be used to help identify spots in a 2D electrophoresis gel. Spots not identified by mass spectrometry are putatively assigned to proteins in the network by comparing their molecular weights and isoelectric points.

Clustering proteins by their GO terms

Networks can become very complex and hence, clustering methods are needed to facilitate their interpretation. PIANA provides a library for applying agglomerative hierarchical clustering to protein interaction networks. For example, using the annotation provided by GO, PIANA groups in clusters those proteins in the network that have similar biological processes or molecular functions. The distance function used for this clustering is based on the length of the path between the GO terms in the GO hierarchical tree. The stop condition is set by the user by means of two thresholds: minimum similarity accepted in order to group two clusters and minimum distance from the terms in the cluster to the GO root term.

Extending PIANA

New functionalities can be added to PIANA by extending the current python classes. Moreover, PIANA implements a class called PianaApi that can be used from other python programs to work with interaction networks.

4 EXAMPLE

We illustrate the use of PIANA with two genes (MMP1 and LTBP1) that have been found to mediate breast cancer metastasis to lung (Minn, et al., 2005). First of all, we create a PIANA configuration file where we set 1) the input parameters (eg. input proteins and network depth), 2) the output parameters (eg. type of protein identifiers to be used) and 3) the PIANA commands to execute (eg. create network for the proteins and predict interactions based on interologs). Then, we run PIANA with the configuration file as an argument. Figure 2.1.1 shows the protein interaction network for MMP1 and LTBP1 before (a) and after (b) adding predictions based on interologs. A detailed PIANA example using all the genes from (Minn, et al., 2005) and performing an in-depth analysis of the interaction network can be found at http://sbi.imim.es/piana/example.html.

Furthermore, PIANA has been previously used for the study of biological pathways in breast cancer cells (Espana et al., 2005).

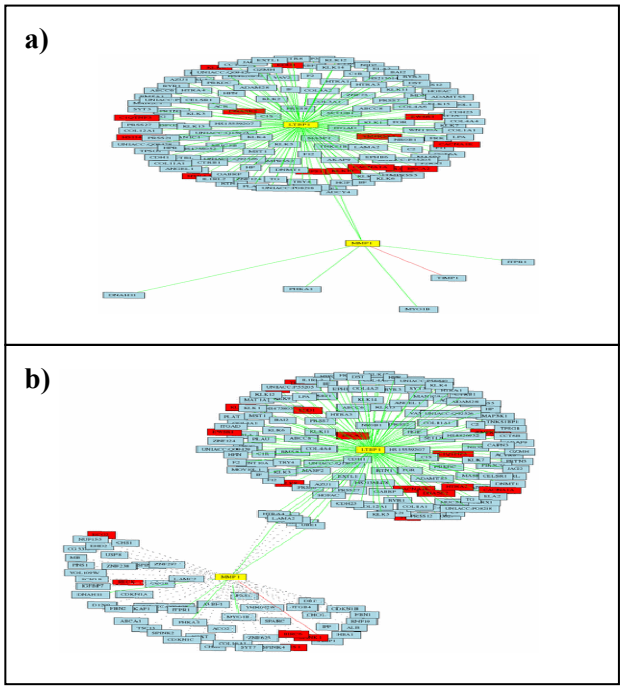


Figure 2.1.1: a) protein interaction network for MMP1 and LTBP1 and **b)** network obtained after adding predictions based on interologs

5 FUTURE WORK

Future plans for PIANA include the annotation of proteins based on network motifs, prediction of protein structure using interactions (Espadaler, et al., 2005) and developing a reliability score for interactions. We intend as well to introduce algorithms that split proteins into the domains that perform the interactions.

ACKNOWLEDGEMENTS

We thank J. Planas, P. Boixeda, B. Gregori and L. Salwinski for their helpful comments. R.A is supported by a grant from the Spanish Ministerio de Ciencia y Tecnología (MCyT, BIO2002-03609). This work has been supported by grants from Fundación Ramón Areces, from the Spanish Ministerio de Educación y Ciencia (MEC, BIO02005-00533), and the ‘‘Programa Gaspar de Portolà (DURSI)’’

REFERENCES

Alfarano, C. et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Res*, **33**, D418-424.

Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bairoch, A. et al. (2005) The Universal Protein Resource (UniProt), *Nucleic Acids Res*, **33**, D154-159.

Benson, D.A. et al. (2005) GenBank, *Nucleic Acids Res*, **33**, D34-38.

Breitkreutz, B.J. et al. (2003) Osprey: a network visualization system, *Genome Biol*, **4**, R22.

Espadaler, J. et al. (2005) Detecting remotely related proteins by their interactions and sequence similarity, *Proc Natl Acad Sci U S A*, **102**, 7151-7156.

Espana, L. et al (2005) Bcl-x(L)-mediated changes in metabolic pathways of breast cancer cells: from survival in the blood stream to organ-specific metastasis, *Am J Pathol*, **167**, 1125-1137.

Hermjakob, H. et al. (2004) The HUPO PSI's molecular interaction format, *Nat Biotechnol*, **22**, 177-183.

Hu, Z. et al. (2005) VisANT: data-integrating visual framework for biological networks and modules, *Nucleic Acids Res*, **33**, W352-357.

Huynen, M.A. et al. (2003) Function prediction and protein networks, *Curr Opin Cell Biol*, **15**, 191-198.

Iragne, F. et al. (2005) ProViz: protein interaction visualization and exploration, *Bioinformatics*, **21**, 272-274.

Minn, A.J. et al. (2005) Genes that mediate breast cancer metastasis to lung, *Nature*, **436**, 518-524.

Murzin, A.G. et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, **247**, 536-540.

Pagel, P. et al. (2005) The MIPS mammalian protein-protein interaction database, *Bioinformatics*, **21**, 832-834.

Peri, S. et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Res*, **13**, 2363-2371.

Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions, *Curr Opin Struct Biol*, **13**, 377-382.

Salwinski, L. et al. (2004) The Database of Interacting Proteins, *Nucleic Acids Res*, **32**, D449-451.

Shannon, P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res*, **13**, 2498-2504.

Tatusov, R.L. et al. (2003) The COG database, *BMC Bioinformatics*, **4**, 41.

von Mering, C. et al. (2003) STRING: a database of predicted functional associations between proteins, *Nucleic Acids Res*, **31**, 258-261.

Yook, S.H. et al. (2004) Functional and topological characterization of protein interaction networks, *Proteomics*, **4**, 928-942.

Yu, H. et al. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs, *Genome Res*, **14**, 1107-1118.

Assessment of protein-protein interaction data in the public domain by integration of multiple sources

Ramón Aragues[1*], Javier García-García[1*], Baldo Oliva[1, §]

1. Structural Bioinformatics Lab. (GRIB). Universitat Pompeu Fabra-IMIM. Barcelona Research Park of Biomedicine (PRBB). 08003-Barcelona, Catalonia, Spain.

[§] Corresponding author [*] Both authors contributed equally to this work

Email addresses: RA: ramon.aragues@upf.edu; JG: jgarcia1@imim.es; BO: boliva@imim.es

Abstract

Background. The analysis and usage of biological data is hindered by the spread of information across multiple repositories and the difficulties posed by different nomenclature systems and storage formats. In particular, the study and use of protein-protein interactions is one area where there is an important need for data integration and improvements on the tools used to gather information. Without good integration strategies, it is difficult to assess how much interaction data is available and its properties.

Results. We present a data integration approach for protein-protein interactions. Our approach is different to other techniques in that 1) it uses the sequence of the protein and its taxonomy as the unique identifier; 2) it integrates most protein and interaction repositories into a single relational database; and 3) the integrated protein interaction network can be built for any source database, detection method and species, or combinations of them. This integrative approach has been implemented into PIANA, a protein-protein interaction software framework under the GNU Public License (http://sbi.imim.es/piana). The low overlap between the different sources of interaction data highlights the need for integrative tools. We find that the integrated network of interactions shows properties very similar to those previously observed in previously reported protein interaction networks.

Conclusions. PIANA´s approach to protein interaction data integration solves many of the nomenclature issues common to systems dealing with biological data. The concept presented here can be extended to other types of biological data. The integration of all available protein interaction data is fundamental to obtaining a comprehensive picture of the interactions taking place in the cell.

Background

The completion of genome sequencing projects stimulated the development of high-throughput experimental methods aimed at functional characterization of the discovered genes. In particular, the identification of protein-protein interactions has been accelerated by the development of new technologies such as two-hybrid assays [1-3] and affinity purifications followed by mass spectrometry [4-6]. Thus, a vast amount of protein-protein interaction data has been collected, including proteome-scale interactome maps for yeast [7, 8], fly [9] and worm [10], and a partial map for human [2, 3]. In addition to providing insights about biological systems [11, 12], protein interaction maps can be used to infer the function of proteins (ref), detect remote homologs (ref) and to identify the interacting motifs of a protein (ref).

However, interaction data is spread across multiple repositories and codified using various protein nomenclature systems [13]. In consequence, experimental biologists face difficulties when trying to find all known interactions for their protein of interest, and the computational analysis and use of protein interaction data is usually restrained to a partial subset of all available knowledge. For example, any comprehensive search of interactions for a particular protein must include at least seven databases of protein-protein interactions: the Database of Interacting Proteins (DIP) [14], the MIPS database of interactions [15], the Molecular INTerations database (MINT) [16], IntAct [17], the Biomolecular Interactions Database (BIND) [18], the BioGrid interactions database (ref) and Human Protein Reference Database of interactions (HPRD) [19]. Besides, each database uses different strategies for identifying proteins, and translations between synonyms are required before any manual search or automatic processing. Moreover, there are methods for predicting protein interactions that can be used when no experimental interactions have been detected for a protein, but results from these methods are usually spread across multiple websites, each one in its own format flavor. Although there have been efforts to standardize the format for protein-protein interactions data exchange [20], the guidelines implemented do not include a strategy for naming proteins, which leaves unresolved many of the integration issues.

The importance of protein interactions analysis has prompted the development of tools focused on protein interaction networks and their visualization, analysis and data integration [21]. For example, Cytoscape is a major effort to centralize network analysis tools on a single platform with built-in visualization [22]. Other visualization and analysis tools

include Osprey (ref), VisANT [23], and ProViz [24]. On the other hand, current packages aimed at data integration include tYNA [25], a web system for managing, comparing and mining multiple networks, and cPath [26], a platform for collecting, storing and biological pathways that can be used from third party softwares for visualization and analysis. While these tools have been shown to be useful for creating and analyzing protein-protein interaction networks, there is still the need for an integration engine that truly unifies all available data into a single network and allows automatic analyses on a global scale. Most current integration tools are designed to work with interactions coming from one single type of data format, and others have problems when dealing with interactions codified using different types of protein identifiers.

Recently, a number of studies have examined the protein interaction data available in the public domain [13, 27, 28]. Pandey and coworkers [13] analyzed human experimentally detected interactions from multiple databases, concluding that repositories show little overlap among them. Herzel *et al*. [28] also compared human interaction maps, but added interaction predictions to the list of analyzed repositories. They concluded that the overlap between repositories is small but significant, and showed strong sampling and detection biases could be linked to the different interacion maps. The integration strategy of both works consisted in mapping all binary interactions to pairs of Entrez Gene identifiers. Marcotte and coworkers [27] analyzed yeast and human interaction data sets, and estimated their protein interaction networks to contain 37,800-75,500 and 154,000-369,000 interactions respectively.

In a recent work, we presented PIANA (Protein Interactions And Network Analysis), a framework for creating, managing and analyzing protein-protein interactions [29]. Here, we describe the PIANA approach to protein nomenclature and its strategy to protein-protein interaction data integration. Furthermore, we describe the properties of the interaction network obtained for all species by integrating interactions from DIP [14], MIPS [15], MINT [16], IntAct [17], BIND [18], BioGrid (ref) and HPRD [19]. We also describe the properties of the interaction networks obtained from different methods for predicting protein interactions. We conclude by discussing potential enhancements to the integration approach here described.

Results

Overview

PIANA (Protein Interactions And Network Analysis) [29] is a software framework (Figure 1) capable of (i) integrating multiple sources of information into a single relational database (Supplementary Figure 1); (ii) creating and analyzing protein interaction networks; and (iii) mapping multiple types of biological data onto protein interaction networks. PIANA code and documentation are freely available under an open source license for local installation and modification (http://sbi.imim.es/piana).

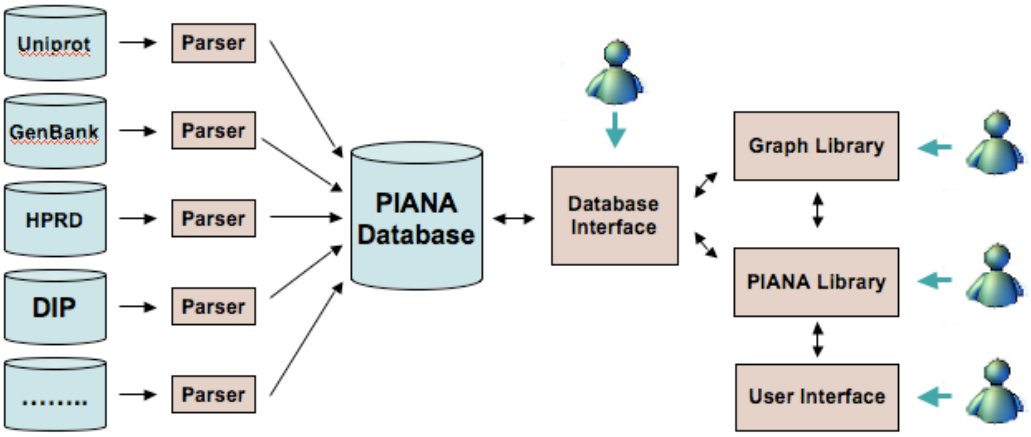


Figure 1: PIANA architecture: a set a parsers inserts information from external repositories into the PIANA database. This database is accessed through an interface by the Graph objects, which is used by the PIANA library to create, manage and analyze protein-protein interaction networks. The whole process can be controlled from a user interface module.

PIANA and protein identifiers

PIANA 'understands' an extensive set of protein identifiers types: UniProt entries and accessions; gene symbols; NCBI gi, geneID, Unigene and accession numbers; ENSEMBL; RefSeq; PDB; and FastA formatted sequences. PIANA internally identifies proteins with proteinIDs (integers). Each proteinID is linked to a pair [aminoacid sequence, taxonomy id], so there is a unique identifier for each protein sequence for a given organism. This allows PIANA to use the sequence of the protein as an inter-lingua between the external identifiers provided by the main repositories of genes and proteins. Therefore, one external protein identifier (e.g. UniProt entry THRB_HUMAN) can be associated to one or more proteinIDs (e.g. 11483), which are in turn linked to other external identifiers that are also used to

represent that protein (e.g., gene symbol ‘f2’ and Unigene ‘Hs.410092’). Therefore, at any point in the different processes involved in inputting/outputting PIANA (e.g. printing the interaction network), external identifiers are ‘translated’ to proteinIDs, the desired operations are performed, and finally, if needed, proteinIDs are ‘translated back’ into the external identifier expected by the user (Figure 2). This strategy reduces the ambiguity and processing problems to the minimum: there is no need for continuously translating between protein identifiers synonyms, since all information has been previously stored by assigning it to specific proteinIDs.

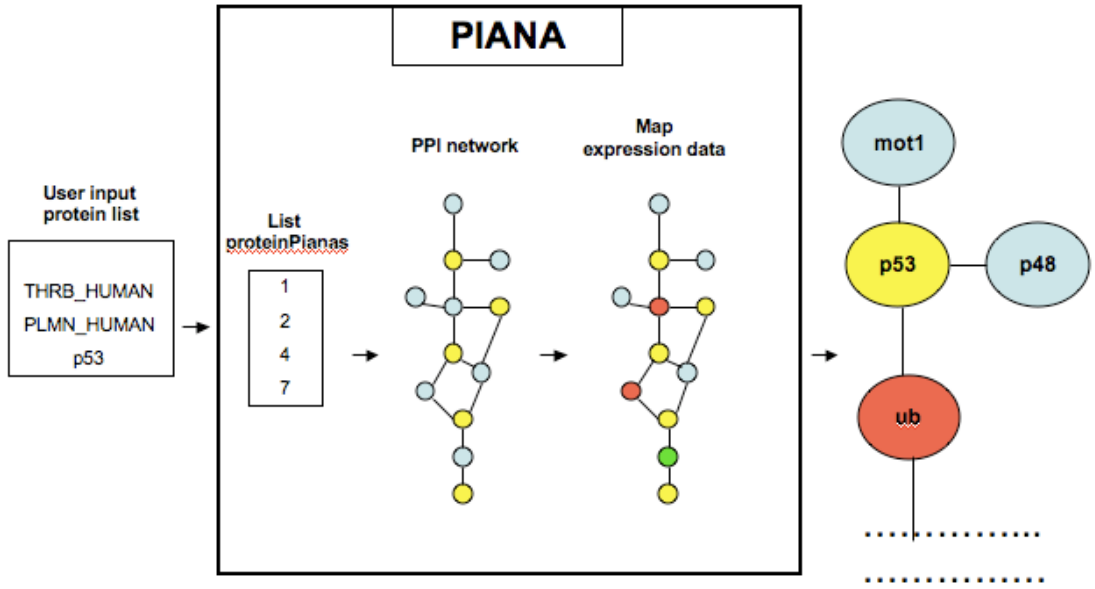


Figure 2: PIANA analyses involve translating from/to proteinIDs. Once this translation has been performed, all operations are performed at the sequence level, reducing ambiguities and synonyms conversions to a minimum.

Moreover, PIANA uses a number of techniques to assure the quality and completeness of the identifiers used as input/output: 1) self inference on the correspondence between identifiers and sequences even in the case that no external database explicitly contained the cross-reference; 2) uniqueness of output protein identifiers: if two proteinIDs have a common value for the type of identifier demanded by the user, those proteins are considered to be the same for that specific output, and hence, merged into a single network node; 3) avoiding gene name ambiguities: thanks to integrating the species of the protein into our internal identifier, gene names are not confounded even if the same identifier is used for several species; and 4) using representative protein identifiers: PIANA will use the identifier

labeled as “preferred” by the source database (eg. official gene symbol) unless the user says the contrary; any input identifiers given by the user are prioritized over other identifiers in the PIANA database.

Protein sequences integration

Sequence and taxonomy data was obtained from UniProt [30], NCBI GenBank [31] and NCBI Blast nr [32] databases (see additional file 1). Interestingly, UniProt swissprot (i.e. curated sequences) and UniProt trembl (predicted sequences) have a high unexpected overlap (see additional file 2). Moreover, the overlap between Trembl and GenBank is lower than anticipated. Cross-references between external identifiers and proteinIDs were obtained from multiple third-party repositories (see additional file 3). Table 1 shows the coverage provided by the main external identifiers for all proteinIDs in the PIANA database.

Table 1: Protein Identifiers statistics. Summary of the most relevant protein identifier types, calculated from a total of 6,476,028 distinct sequenceIDs in the database. Columns are: Identifier type, number of distinct identifiers, the proportion of proteinID with respect to external identifier correspondences, the proportion of external identifiers with respect to proteinIDs, and the percentage of proteinIDs covered by the external identifier.

Identifier Type	Number of distinct identifiers	External Identifier: Piana proteinID		Piana proteinID : External Identifier		% sequences coverage
Uniprot Accession	4,639,397	1:1	99.74	1:1	96.98	67,81
		1:>2	0.07	1:>2	0.98	
NCBI Accession	10,760,685	1:1	98.44	1:1	24.85	95,53
		1:>2	0.83	1:>2	26.78	
NCBI geneID	2,416,561	1:1	90.77	1:1	98.41	42,29
		1:>2	1.87	1:>2	0.17	
Gene Symbol	4,143,090	1:1	81.10	1:1	79.58	79,53
		1:>2	3.95	1:>2	6.57	
Primary Gene Symbol	1,207,358	1:1	90.34	1:1	97.70	41,79
		1:>2	5.27	1:>2	0.13	

Protein-protein interactions integration

Each interaction described in a third-party database is ‘translated’ to one or more interactions between proteinIDs. For example, if the external database contains an interaction between proteins A and B, with A corresponding to two proteinIDs (e.g 1 and 2)

and B to one proteinID (e.g. 3), two interactions (1-3 and 2-3) will be inserted into the PIANA database. Both interactions will be described in the PIANA database as coming from that specific external database and labeled with the method used to detect the interaction between A and B. This methodology allows PIANA to give full control to the user: 1) interactions can be retrieved from any type of identifier; 2) a network can be created for a given external database (e.g. use only interactions from IntAct) and/or a specific method (e.g. do not use interactions detected in two hybrids assays) and/or a species (e.g. only interested in human interactions); 3) PIANA outputs can be set to use any type of protein identifier and therefore, interactions between proteinIDs are transformed to non-redundant interactions between protein identifiers (Material and Methods). Moreover, relating interactions to protein sequences instead of external identifiers provides a true integration of all known interactions into a single network, while keeping record as well of the source databases and detection methods associated to the interactions. Currently, PIANA can integrate interactions from DIP [14], MIPS [15], MINT [16], IntAct [17], BIND [18], BioGrid, HPRD [19], STRING [33], interactions predicted by distant conservation of sequence patterns and structure relationships [34], interactions transferred between proteins based on orthology [35] and, in general, any interaction data that is in tabulated or PSI-MI [20] formats.

PIANA parameters and commands: update

Most PIANA commands and parameters are described elsewhere [29]. In addition to the new data integration features here described, capabilities recently added to the platform include: 1) mapping known pathways and disease genes to the protein interaction network; 2) assigning interacting motifs to proteins of the network using a clustering algorithm; 3) mapping microarray expression data onto the network; 4) producing output in a format readable by Cytoscape [22]; 5) fine-tuning the network to contain interactions from specific species/databases/detection methods; 6) performing text-mining on the function/description of the network proteins; and 7) saving/loading PIANA sessions.

Experimental interactions

The integrated set of experimental interactions consisted of 4,055,698 interactions between 113,785 different proteinIDs. When grouping proteinIDs by their associated NCBI Gene ID, there were 405,808 interactions for 53,143 proteins. All results presented here are for the interactions grouped by GeneID (Material and Methods). Gene IDs were chosen because,

although they only covered 42% of proteinIDs in the database, the cardinality proteinPiana:geneID was the highest (Table 1).

Interactions distribution

The experimental interactions in the PIANA database have been obtained from 7 different databases, belong to 736 different species, and were detected using 106 different experimental methods. The species with the largest number of experimental interactions are yeast (111,535 interactions) and human (110,457 interactions) (Table 2).

Table 2: Number of interactions by species.

Species	#interactions	#proteins
111535	Saccharomyces cerevisiae	6493
110457	Homo sapiens	36900
90562	Drosophila melanogaster	11605
16097	Escherichia coli	3467
9184	Caenorhabditis elegans	3959
4776	Mus musculus	3495
2723	Plasmodium falciparum 3D7	1279
17021	Other species	9724

Most interactions were found in just one database and were detected by just one method (Figure 3). The high correlation between number of methods for interaction and number of databases is explained by the fact that most interactions appear in just one external repository, and these repositories usually label interactions with a single detection method.

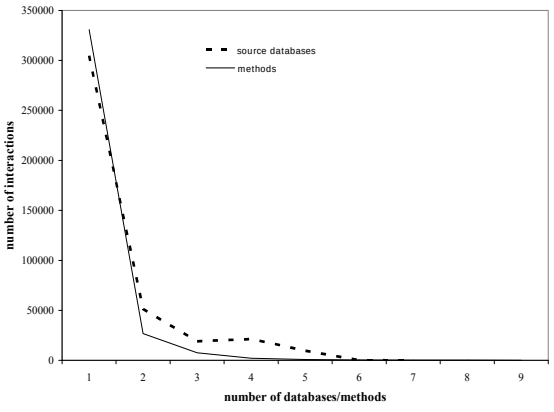


Figure 3: Distribution of interactions in PIANA across different source databases and detection methods. Most interactions were found in just one database and were detected just by one method. Unspecific methods were not taken into account (experimental, invitro, invivo)

We calculated the interactions overlap between the 7 repositories with experimental information that are integrated in PIANA (Table 3). BioGrid is the repository with the highest number of interactions (216,370) and with the highest number of unique interactions (163,700). The two repositories that show the greatest overlap are MINT and IntAct (47,119 interactions). Moreover, we examined the number of interactions detected by the different methods (Table 4). We observed that high-throughput methods account for most of the known interactions.

Table 3: Overlap of protein interactions for the seven databases integrated within PIANA. Pairwise overlaps of protein interactions are shown in cells. For each repository, the total and the unique (interactions only contained in that repository) number of interactions are shown.

Total	Unique							
104339	42284	**IntAct**						
97377	46115	**DIP**	36,867					
77419	16392	**MINT**	47,119	37,210				
38372	10978	**HPRD**	8,729	825	8,925			
833	389	**MIPS**	87	34	107	312		
216370	163700	**BioGrid**	25,355	19,870	24,709	20,550	210	
62444	24925	**BIND**	27,269	28,143	26,406	2,839	121	16,187
			IntAct	**DIP**	**MINT**	**HPRD**	**MIPS**	**BioGrid**	**BIND**

Interaction Detection method	**Number of interactions**
Affinity methods (e.g. TAP)	126,136
yeast two-hybrid assay	103,334
Phenotypic	72,159
3D structure	6,525
Array technologies	4,627
Dosage	3,914
Cross Linking	3,104

Table 4: Number of interactions per detection method. The number of interactions per detection method were calculated after unifying the protein interaction netwok by NCBI geneID and manually grouping similar detection methods.

Properties of the experimental protein interaction network

Well-documented observations about protein interaction networks are confirmed when analyzing the integrated experimental interaction networks of different species. For example, we observed that the networks for the main organisms are are scale free (Table 5) [36, 37]. In addition, the following properties were observed for the yeast protein interaction network (Table 7): (ii) yeast hubs (proteins with >= 5 ints) are more likely to be essential than non-hubs (22% of hubs are essential versus only 5% of non-hubs), although this might be a reflection of hubs usually having multiple interfaces [39]; 2) approximately 59% of the interactions have the same cell localization; 3) approximately 60% of the interactions reported are found coexpressed during the yeast cell cycle.

Organism	**Gamma**	**Scale-free**
Human	1.37	yes
Yeast	1.11	yes
Fruit fly	1.36	yes
E. Coli	1.44	yes
C. Elegans	1.77	yes
Mouse	1.98	yes
H. pylori	1.6	yes
Rat	2.2	not significant
Cow	1.98	not significant

Table 5. Scale-free property for protein interaction networks of the main organisms. For each organism, the gamma value and its significancy are given for the protein interaction network, where gamma is the power-exponent and p-value is the probability that a particular network has such connectivities if they were drawn from the power-law distribution. The scale-freeness was considered significant for p-values below 0.01. The scale-freeness of the interaction networks was calculated using the method by Khanin *et al*. [37].

Table 7: Properties of the yeast protein interaction network obtained by integrating multiple sources with PIANA. Yeast co-localization data was obtained from (ref). Yeast co-expression data was obtained from (ref). Yeast essentiality data was obtained from (ref). A yeast protein was considered a hub if it had 5 or more interaction partners.

	Total number	**Respect property**
Interaction partners are colocalized	72661	37684 (52%)
Interaction partners are co-expressed	2576	1524 (59%)
Yeast hubs	4229	22% essential
Non yeast hubs	886	5% essential

Protein function prediction from the experimental network

Recently, it has been shown that the number of common interaction partners between two proteins can be used to characterize proteins [40, 41]. Here, we have studied the use of this heuristic to predict molecular functions and biological processes as defined by GO (Figure 4). As expected, we observe that the interactions of a protein in the integrated network are an important indication of the function of the protein and the biological processes in which it intervenes.

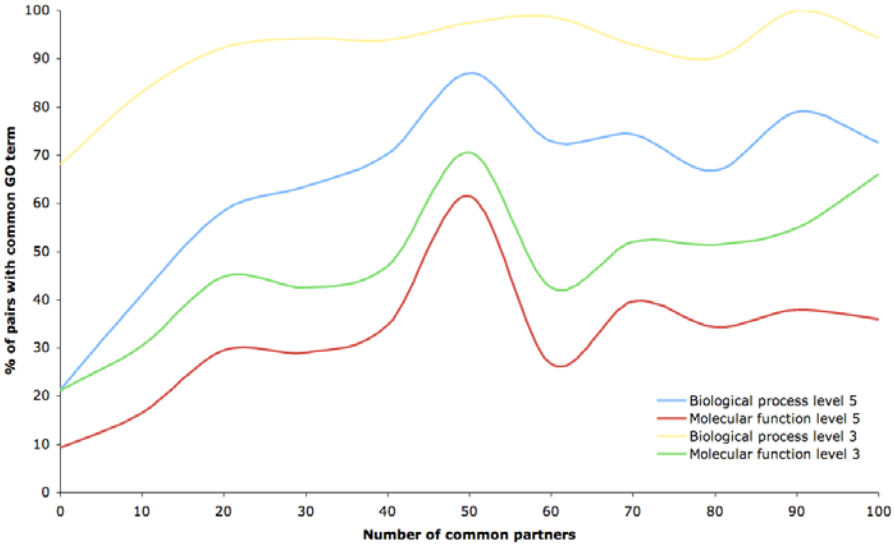


Figure 4. Function prediction based on common interaction partners. The percentage of shared GO terms is plotted as a function of the number of common interaction partners.

Conclusions

We presented the data integration approach of PIANA, a software framework designed for creating, managing and analyzing protein-protein interaction networks. PIANA was created to address nomenclature and integration issues common in protein interaction repositories and network visualization tools. Moreover, the modular approach of PIANA makes it a useful resource for bioinformaticians willing to avoid the low-level details related to working with protein interaction networks. PIANA is one of the very few protein interaction platforms where all interactions from all external databases can be found for a protein of interest, regardless of the type of identifier used as input or the name given to the protein by the researcher that submitted the interactions. We also presented a detailed analysis of the protein-protein interactions included in PIANA, in terms of their distribution across different databases and detection methods. Most importantly, we showed that the overlap between the different repositories is low, which reinforces the need for tools that unify all known interactions into a single network, which then can be used to perform relevant analyses. Moreover, this unified network has been shown to respect properties previously found about protein-protein interaction networks retrieved from just one database/detection method, such as scale-freeness or its use for protein function prediction.

We believe PIANA's approach to data integration is a good equilibrium between reliability and flexibility, while giving a good coverage of the information available. Many areas of biological research are hampered by the difficulty in accessing all biological information available. In particular, protein-protein interactions analysis is usually biased by the input sources of data. We showed that the overlap between the different databases and methods is very low, which reinforces the need of reliable integration systems. The data integration techniques described here could also be of help for areas other than protein-protein interactions, such as gene expression studies or regulatory networks.

Material and Methods

Interaction networks based on proteinIDs or other identifiers

Interaction networks are built using PIANA proteinIDs as nodes (see sections "PIANA and protein identifiers" and "Protein-protein interactions integration"). When translating the network to an external protein identifier (process referred as 'unifying the network'), there are two possibilities: 1) the proteinID corresponds to a single identifier and 2) different

proteinIDs correspond to the same identifier, and thus, nodes and interactions will be merged. Therefore, the same PIANA proteinID network will correspond to different unified networks, depending on the external identifier used to unify. Statistics in this article have been obtained after unifying the networks by NCBI geneID. Although it only covers 42% of all proteinIDs, the cardinality proteinID:externalIdentifier is the highest (Table 1), and therefore it is the best suited identifier type for obtaining an unbiased view of the protein interaction network.

Databases parsing

Most databases of protein-protein interactions do not follow a standardized nomenclature system to describe interaction detection methods. Based on the Open Biomedical Ontologies , we have manually created a controlled vocabulary of detection method names, and all third-party method labels have been internally mapped to the internal method names. When considered necessary, very speficic method names were merge into a more general term (e.g., different enzimatic methods have been joined as "enzimatic method"). The overlap results presented here have been obtained after eliminating those methods considered too generic (e.g. experimental, invivo, invitro).

Authors' contributions

RA designed PIANA and wrote the first draft of this manuscript. JG and RA implemented the code. BO conceived of the PIANA project, provided scientific guidance and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the UPF-IMIM SBI lab and P. Boixeda for their helpful comments. R.A is supported by a grant from the Spanish Ministerio de Ciencia y Tecnología (MCyT, BIO2002-03609). The work has been supported by grants from Fundación Ramón Areces, from the Spanish Ministerio de Educación y Ciencia (MEC, BIO02005-00533), and the ''Programa Gaspar de Portolà (DURSI)''.

References

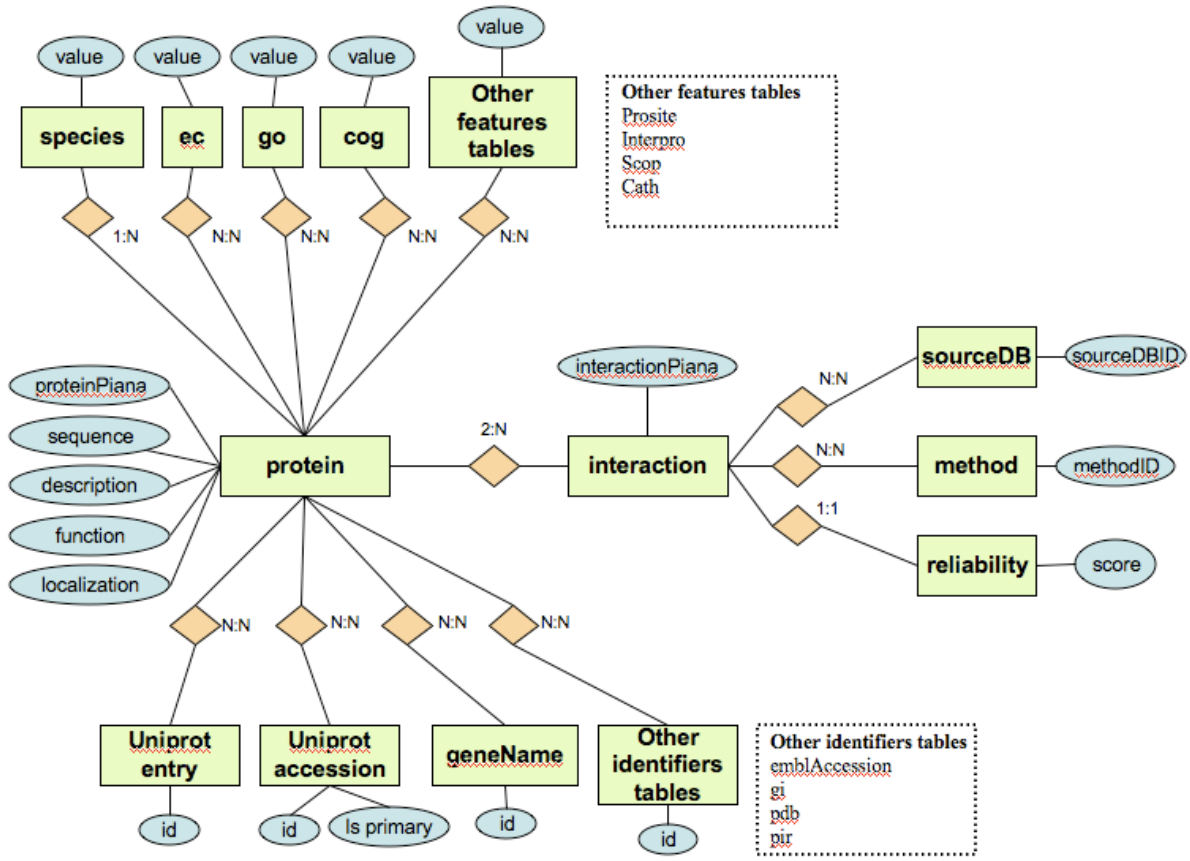
1. Parrish, J.R., K.D. Gulyas, and R.L. Finley, Jr., *Yeast two-hybrid contributions to interactome mapping.* Curr Opin Biotechnol, 2006. **17**(4): p. 387-93.

2. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.
3. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.
4. Puig, O., et al., *The tandem affinity purification (TAP) method: a general procedure of protein complex purification.* Methods, 2001. **24**(3): p. 218-29.
5. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.
6. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.
7. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
8. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.
9. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
10. Li, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
11. Cusick, M.E., et al., *Interactome: gateway into systems biology.* Hum Mol Genet, 2005. **14 Spec No. 2**: p. R171-81.
12. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
13. Mathivanan, S., et al., *An evaluation of human protein-protein interaction data in the public domain.* BMC Bioinformatics, 2006. **7 Suppl 5**: p. S19.
14. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
15. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-4.
16. Chatr-aryamontri, A., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
17. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.
18. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33**(Database issue): p. D418-24.
19. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
20. Hermjakob, H., *The HUPO Proteomics Standards Initiative - Overcoming the Fragmentation of Proteomics Data.* Proteomics, 2006. **6 Suppl 2**: p. 34-8.
21. Aittokallio, T. and B. Schwikowski, *Graph-based methods for analysing networks in cell biology.* Brief Bioinform, 2006. **7**(3): p. 243-55.
22. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
23. Hu, Z., et al., *VisANT: an online visualization and analysis tool for biological interaction data.* BMC Bioinformatics, 2004. **5**: p. 17.
24. Iragne, F., et al., *ProViz: protein interaction visualization and exploration.* Bioinformatics, 2005. **21**(2): p. 272-4.
25. Yip, K.Y., et al., *The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks.* Bioinformatics, 2006. **22**(23): p. 2968-70.

26. Cerami, E.G., et al., *cPath: open source software for collecting, storing, and querying biological pathways.* BMC Bioinformatics, 2006. **7**: p. 497.
27. Hart, G.T., A.K. Ramani, and E.M. Marcotte, *How complete are current yeast and human protein-interaction networks?* Genome Biol, 2006. **7**(11): p. 120.
28. Futschik, M.E., G. Chaurasia, and H. Herzel, *Comparison of human protein-protein interaction maps.* Bioinformatics (Oxford, England), 2007. **23**(5): p. 605-11.
29. Aragues, R., D. Jaeggi, and B. Oliva, *PIANA: protein interactions and network analysis.* Bioinformatics, 2006. **22**(8): p. 1015-7.
30. *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2007. **35**(Database issue): p. D193-7.
31. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.
32. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2007. **35**(Database issue): p. D26-31.
33. von Mering, C., et al., *STRING 7--recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.
34. Espadaler, J., et al., *Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships.* Bioinformatics, 2005. **21**(16): p. 3360-8.
35. Yu, H., et al., *Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.* Genome Res, 2004. **14**(6): p. 1107-18.
36. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-4.
37. Khanin, R. and E. Wit, *How scale-free are biological networks.* Journal of computational biology, 2006. **13**(3): p. 810-8.
38. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
39. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.
40. Brun, C., et al., *Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.* Genome Biol, 2003. **5**(1): p. R6.
41. Samanta, M.P. and S. Liang, *Predicting protein functions from redundancies in large-scale protein interaction networks.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12579-83.

Selected additional files

Additional file 1. The relational design of the PIANA database



CHAPTER III

PROTEIN HUBS CHARACTERIZATION BY INFERRING INTERACTING MOTIFS FROM PROTEIN INTERACTIONS

Proteins interact with other proteins through a limited set of interface types. However, most methods for protein-protein interaction detection do not identify the regions of the proteins that are in contact during the interaction. In this chapter, we include an article submitted to PLoS Computational Biology (in revision) where we investigate the use of PIANA for addressing two interesting questions:

1. Can we use protein-protein interaction data to throw some light over the interfaces of protein?

2. What is it that makes protein hubs different from other proteins in terms of essentiality and evolutionary rate?

We found, similarly to other works based on 3D structures, that it is the number of distinct interfaces of a hub that makes it more essential and evolve slower than other proteins, rather than the traditional explanation based on their high absolute number of interactions

Articles included in this chapter:

Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. **Protein Hubs Characterization by Inferring Interacting Motifs from Protein Interactions.** PLoS Computational Biology (in revision)

Protein Hubs Characterization by Inferring Interacting Motifs from Protein Interactions

Ramon Aragues[1], Andrej Sali[2], Jaume Bonet[1], Marc A. Marti-Renom[3*], and Baldo Oliva[1*]

[1] Structural Bioinformatics Lab. (GRIB). Universitat Pompeu Fabra-IMIM. Barcelona Research Park of Biomedicine (PRBB). 08003-Barcelona, Catalonia, Spain.

[2] Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94158-2330, USA.

[3] Structural Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), 46013-Valencia, Spain.

[*] Corresponding authors

Baldo Oliva

Laboratori de Bioinformàtica Estructural (GRIB), Universitat Pompeu Fabra-IMIM. Parc de Recerca de Biomedicina (PRBB). C/ Doctor Aiguader, 88, Barcelona 08003, Catalonia, Spain. Tel: +34933160509; Fax: +34933160550; e-mail: boliva@imim.es

Marc A. Marti-Renom

Structural Genomics Unit, Bioinformatics Department. Centro de Investigación Príncipe Felipe. Av. Autopista del Saler, 16, 46013 Valencia, Spain. Tel: +34 96 3289680 Fax: +34 96 3289701. e-mail: mmarti@cipf.es

Abbreviations: iMotif: interacting motif;

Abstract

Characterization of protein interactions is essential for understanding biological systems. While genome-scale methods are available for identifying interacting proteins, they do not pinpoint the interacting motifs (e.g., a domain, sequence segments, a binding site, or a set of residues). Here, we develop and apply a method for delineating the interacting motifs of hub proteins (i.e., highly connected proteins in the interactome) by relying on the observation that proteins with common interaction partners tend to interact with these partners through a common interacting motif. The sole input for the method are binary protein interactions; neither sequence nor structure information is needed. The approach is evaluated by comparing the inferred interacting motifs to domain families defined for 368 proteins in the Structural Classification of Proteins (SCOP). The positive predictive value of the method in

detecting proteins with common SCOP families is 75% at sensitivity of 10%. We find that yeast hubs with multiple interacting motifs are more likely to be essential than hubs with one or two interacting motifs, thus rationalizing the previously observed correlation between essentiality and the number of interacting partners of a protein. We also find that yeast hubs with multiple interacting motifs evolve slower than the average protein, contrary to the hubs with one or two interacting motifs. The proposed method will help discover unknown interacting motifs and provide biological insights about protein hubs and their roles in interaction networks.

Introduction

Protein-protein interactions play a central role in many cellular processes, ranging from signal transduction to formation of cellular macrostructures and cell cycle control [1-3]. Recently, several techniques such as two-hybrid assays [4-6] and affinity purifications followed by mass spectrometry [7-9], have enabled large-scale identification of protein-protein interactions. While these efforts provide rich lists of interacting proteins, they do not produce information about the specific interfaces involved in each interaction.

Proteins interact through a limited set of interface types [3,10,11]. The interfaces are usually key determinants of the function. Therefore, narrowing down protein-protein interactions to interactions between specific protein components (e.g., a domain, sequence segments, a binding site, or a set of residues) is important for a more accurate characterization of the function of proteins and their complexes. Identifying the protein interfaces that mediate interactions may also be useful for the prediction of unknown protein-protein interactions [12], for homology-based protein annotation methods [13], and for relating gene essentiality and network topology [14].

Traditionally, the description of protein interactions in terms of the interacting components has been based on protein structural domains [15], protein functional sites [16], and protein patches [17]. However, fully characterizing protein surfaces that are in contact with each other during an interaction requires the determination of the structure of protein complexes by X-ray crystallography or NMR spectroscopy. These methods are not always applicable and thus the number of known 3D atomic structures of proteins and their complexes is limited. As a result, accurate and general computational methods for identifying motifs involved in protein-protein interactions are needed.

Recently, several methods [18-21] have been developed to describe protein-protein interactions in terms of interacting protein domains, as defined in the Structural Classification of Proteins (SCOP) [22], PFAM [23], and InterPro [24] databases. However, while these methods find interactions between predefined protein domains, interactions between undefined domains remain undetected. Structure-based methods overcome this problem by predicting the amino acid residues that are in contact during a protein-protein interaction, but require the structures of both proteins [25-28]. Recently, Kim *et al.* used known protein interactions and structures to characterize the interfaces between two interacting proteins [14]. They found that some previously accepted relationships between network topology and genomic features [29-31] are actually more reflective of the number of distinct binding interfaces. For example, highly connected proteins in the network (i.e., hubs) with multiple interfaces are twice as likely to be essential as hubs with one or two interfaces. The findings of Kim and coworkers clarify some previous analyses that related the observed essentiality of hubs with their high number of interacting partners [29,32] or with their interactions to other hubs [33]. Kim *et al.* also demonstrated that the evolutionary rate is significantly lower for multi-interface hubs than for the average protein, but not so for hubs with one or two interfaces.

Here, our basic assumption is that proteins with overlapping sets of interaction partners tend to interact with the common partners through the same interacting motif, such as a domain, sequence segments, a binding site, or a set of residues. A similar assumption has been previously used to annotate protein sequences [13,34-36]. We first tested this assumption based on databases of protein interactions [37] and protein domains defined in SCOP [22], observing that the assumption holds true for highly connected proteins (i.e., hubs). Building on this validation, we then developed a method for identifying interacting motifs (iMotifs) in hub proteins, which has been implemented within the protein-protein interaction framework and integration engine PIANA (Protein Interactions And Network Analysis) [37]. iMotifs are not required to be of any particular structural type or size, thus allowing us to characterize hub proteins and their interactions at different levels of resolution, ranging from full proteins to small binding sites. In contrast to other methods, our approach is not limited to finding predefined classes of interacting motifs, such as SCOP domains or PROSITE functional sites, and can be used to identify unknown interacting motifs. Moreover, the sole input for our method are binary protein interactions; neither structure nor sequence information is required to assign iMotifs to proteins.

Two main objectives have been addressed in this work. The first objective was to demonstrate whether protein interactions alone can be used to infer interacting motifs. The positive predictive value of our method in detecting proteins with common SCOP families was 75% at sensitivity of 10%, and the Spearman correlation coefficient between the number of iMotifs assigned to proteins and the number of interfaces found by Kim *et al.* [14] was 0.57. The second objective was to examine if the conclusions on protein hubs of Kim *et al.* [14] hold for our iMotifs assignments. The results demonstrate that protein hubs with multiple iMotifs are more likely to be essential than hubs with one or two iMotifs and that protein hubs with multiple iMotifs evolve slower than the average protein in the dataset, as opposed to hubs with one or two iMotifs.

Results

Proteins with common interaction partners tend to share a SCOP domain

The basic assumption behind this work is that proteins with overlapping sets of interaction partners tend to interact with those partners through a common interacting motif. The validity of this assumption was tested on a nonredundant set of 368 proteins (Material and Methods) by analyzing the relationship between the number of interaction partners shared by two proteins and the likelihood of those proteins having a domain within the same SCOP family [22]. Although SCOP does not classify proteins by their kinds of interfaces, it has been shown that protein interaction types can be defined by the domains in the interacting proteins [38] and thus, in this validation we used SCOP domains as a surrogate for interacting motifs.

We found the number of common interaction partners (N) to be a good indicator of the probability of two proteins having a domain within the same SCOP family, especially for highly connected proteins (Figure S1). For example, 73% of protein pairs with 50-60 common interaction partners shared a SCOP domain. We also studied other metrics to measure the similarity between two sets of interaction partners, but none of them outperformed N at the identification of protein pairs with a common domain family (Figure S1, S2A and Table S1).

Our assumption relies on the binary nature of the interactions used, as two proteins will tend to interact through a common interacting motif only if they have direct physical interactions with the same partners. We examined the effect of restricting the study to interactions detected with the yeast two hybrid method, which is the best suited assay for detecting

binary protein interactions [39]. We did not observe the expected increase of protein pairs with a common domain family (Figure S2B).

Delineating interacting motifs

Based on the observation that highly connected proteins with common interaction partners tend to interact with these partners through a common interacting motif, we have developed a method that groups proteins with similar interacting motifs (Figure 1). The procedure (Methods) is carried out in four steps: 1) build the protein-protein interaction network; 2) initialize the cluster interaction network by assigning each protein of the network to a cluster; 3) iteratively fuse similar clusters (allowing a protein to be in more than one cluster) until the similarity score drops below a predefined threshold; and 4) label with a different interacting motif identifier (iMotif) each cluster with more than one protein, and derive iMotif assignments and iMotif-iMotif interactions from the clustered network.

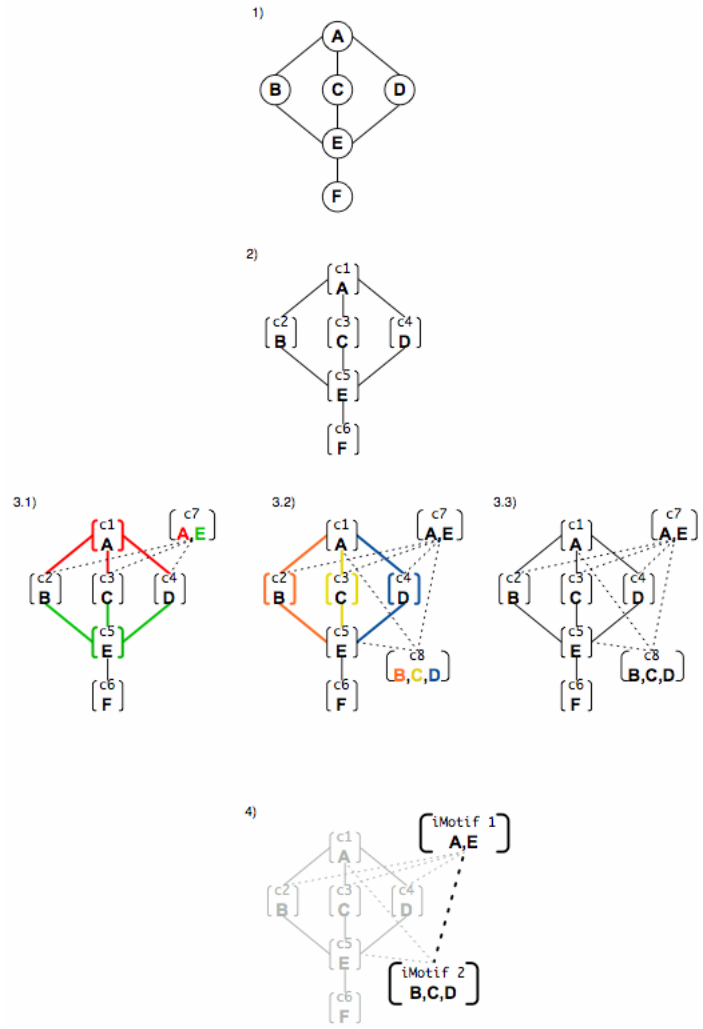


Figure 1. Description of the procedure followed for assigning iMotifs to proteins and identifying iMotif-iMotif interactions. First, the protein interaction network is built. Second, a cluster interaction network is created by placing each protein in a different cluster. Third, clustering is performed until the similarity score drops below a certain threshold. Fourth, an iMotif label is assigned to each cluster with more than one protein, and iMotif assignments and interactions are derived.

Assigning an interacting motif identifier (iMotif) to a group of proteins simply establishes that they have a certain feature that allows them to interact with the same set of partners, without giving information on the size, sequence or structure of that feature (Figure 2A); an iMotif can be an interface consisting of a set of domains or only a specific constellation of a small number of residues (Figure 2B).

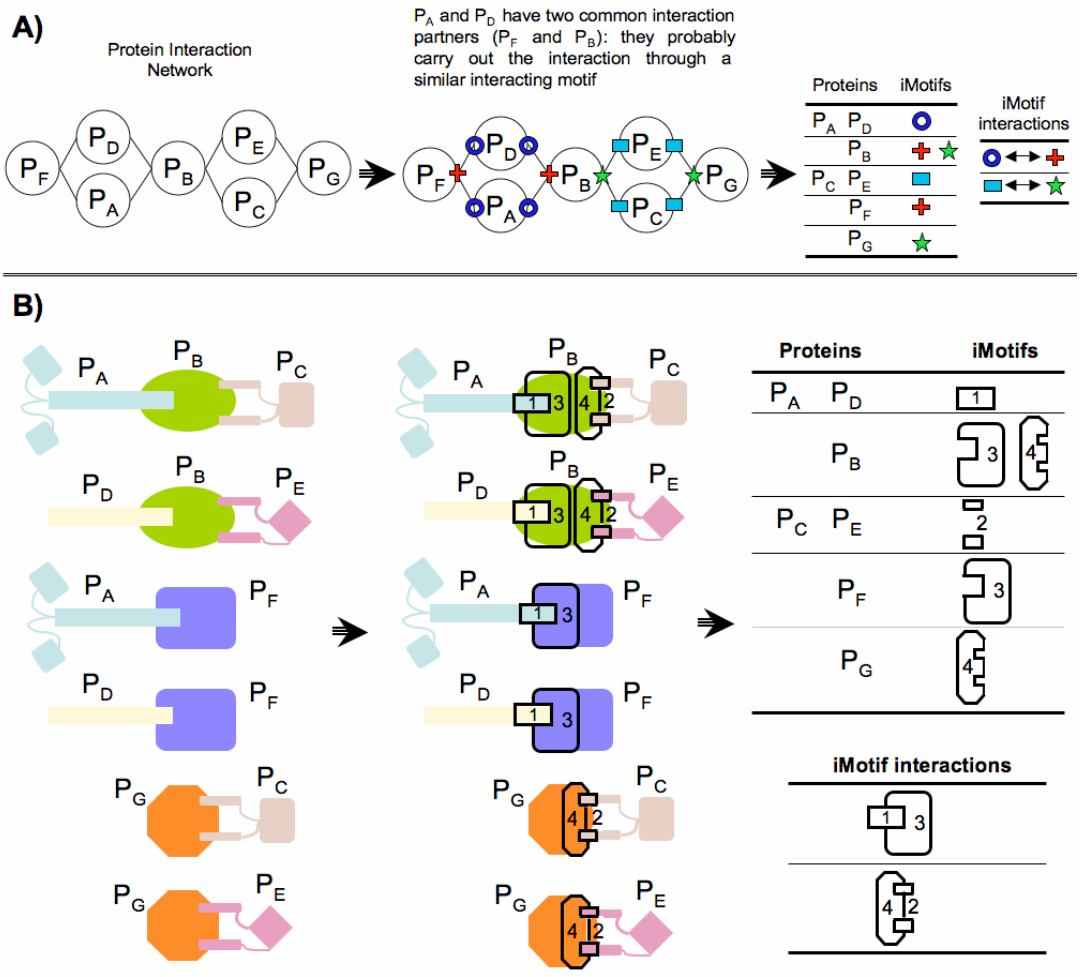


Figure 2. **Definition of an Interacting Motif (iMotif).** The definition of an iMotif depends on the minimum number of common partners required in order to consider the given binary protein interactions mediated through a common interacting motif. A) From the protein interaction network perspective, proteins with common partners (two in the example provided) are considered to interact with these partners through a similar feature, and therefore, are classified as being of the same iMotif. B) The same process is shown from a structural perspective: proteins interacting through a similar feature (regardless of the feature being two structural domains or a single binding site) are considered to have a common iMotif. To further illustrate the method, we also describe a sample iMotif assignment for *prothrombin* (UniProt code THRB_HUMAN) (Text S1 and Figure S3).

Method evaluation

The definition of iMotifs depends on a similarity metric and its threshold. Thus, different thresholds or metrics produce different iMotifs, corresponding to different levels of resolution in the description of protein-protein interactions. For example, the method can be applied at the resolution of domains from SCOP [22], and PFAM [23], or at the higher resolution of functional sites from PROSITE [40]. In this work, we have evaluated the method on a nonredundant set of proteins (Material and Methods) for three different tasks: (i) detecting proteins with common SCOP domain families; (ii) predicting SCOP domain-domain interactions observed in the PDB [41]; and (iii) predicting the number of distinct binding interfaces as defined by Kim *et al.* [14]. Therefore, in the evaluation, iMotifs effectively represent SCOP family domains (for the first two tasks) and binding interfaces (for the third task).

Detecting proteins with domains in the same SCOP family

We evaluated the ability of the method to detect proteins with a domain in the same SCOP family (Methods). Using a threshold of 30 common interaction partners (N), our method achieves a positive predictive value of ~75%, sensitivity of ~10%, and applicability of ~20% (Figure 3). The positive predictive value is above 50% for N thresholds higher than 15 and thus, the method should be preferentially applied to highly connected proteins. While the current utility of the method is limited by its relatively low sensitivity and applicability, the growth of the interactome data [42,43] is likely to make the approach more applicable in the future. Moreover, the applicability can already be increased at the expense of lower positive predictive value by using other similarity metrics (Table S1). We provide a complete list of iMotif assignments for the test set (Table S2).

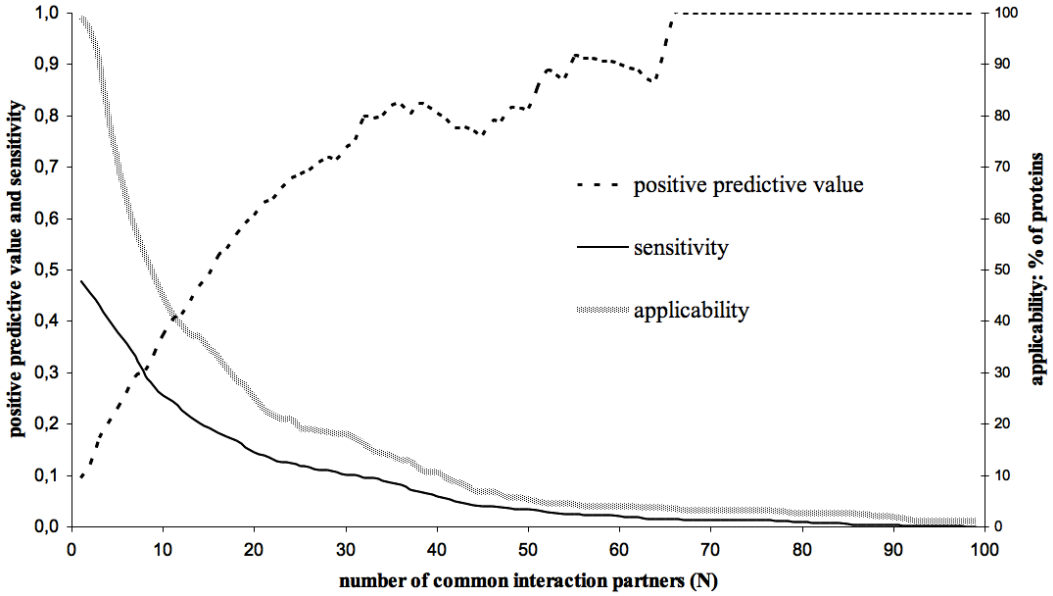


Figure 3. Performance of the method in detecting proteins with common SCOP families. The positive predictive value, sensitivity and applicability (Methods) are plotted as a function of the number of common interaction partners threshold (*N*) used for the clustering. We observe that the method obtains high positive predictive values in detecting proteins with the same SCOP domain when high numbers of interaction partners are shared.

Predicting domain-domain interactions

Domain-domain interactions can be predicted from the iMotif-iMotif interactions found by the method (Methods). We evaluated the accuracy of these predictions with respect to domain interactions in the PDB. Our method achieves a positive predictive value of ~65% for ~5% of the proteins in the test set (Figure S4), suggesting that the method can be applied to the prediction of domain-domain interactions when a sufficiently large and varied sample of protein interactions is known. However, with the amount of data currently available, methods based on interaction networks and predefined domains [18-21] are better suited than our approach for predicting domain-domain interactions.

Predicting the number of binding interfaces

Kim *et al.* used protein 3D structures and binary protein interactions to make inferences about the number of binding interfaces of proteins [14]. We tested whether there is a correlation between the number of binding interfaces found in their work and the number of iMotifs predicted by our method (Figure 4). The number of protein interfaces indeed correlates with the number of predicted iMotifs per protein (e.g., for *N* of 20, r_s is 0.57 and

p-value 0.01). This correlation is significant for all *N* values lower than 23 (Figure S5). We observe that, although correlated, the number of iMotifs assigned to proteins tends to be much higher than the number of binding sites defined by Kim *et al.* [14]. This might be attributed to two factors: (i) an underestimation of the number of binding sites assigned by the method in [14], attributable to the fact that current structural data do not contain all possible protein-protein interactions; and (ii) an overestimation of our method in the number of iMotifs per protein, attributable to lack of coverage of the interactome space.

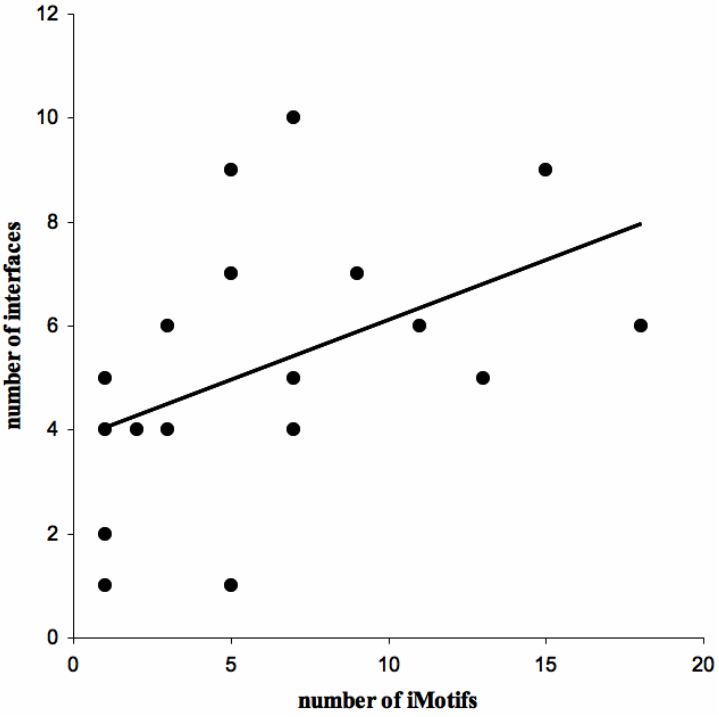


Figure 4. Correlation between the number of binding interfaces and the number of iMotifs. Each point corresponds to a protein from the test set for which a number of binding interfaces was assigned by Kim *et al.* [14] and a number of iMotifs was inferred with N set to 20. Both variables were found to be significantly correlated (r_s is 0.57 and p-value is 0.01). The correlation between the number of interfaces and the number of iMotifs is significant for all *N* values lower than 23 (Figure S5).

iMotif assignments for hub proteins

We applied the method using an *N* threshold of 20 to the 5,571 hubs (i.e., proteins with 20 or more interaction partners) in PIANA. The method assigned 17,403 iMotifs to 2,014 hubs, an average of 8.64 iMotifs per hub. The percentage of hubs with one or two iMotifs was 46% (241 hubs had one iMotif; 689 hubs had two iMotifs). In contrast, the average number of interactions per hub was ∼49. Moreover, we studied the correlation between the number of iMotifs assigned to a hub and its number of interactions, finding no relationship between the two variables. We provide the complete list of iMotif assignments for all hub proteins in PIANA (Data S1) and a simplified table with the number of iMotifs per hub (Table S3).

Essentiality and number of iMotifs are correlated in hub proteins

Similarly to the results by Kim and co-workers [14], we found that yeast hubs with multiple iMotifs are more likely to be essential than those with one or two iMotifs (singlish-iMotif) (Table 1). Furthermore, we observed a correlation (r_s is 0.61 and *p*-value is 1.64×10^{-5}) between the number of iMotifs in yeast hubs and the fraction of essential proteins (Figure 5A). We compared the correlation between iMotifs and essentiality to the correlation between the number of interactions of hubs and essentiality to confirm that the first was not a direct consequence of the second (Figure 5B). These results suggest that the number of iMotifs predicted for a protein could be used for selecting biologically relevant candidates for gene deletion experiments.

Table 1. **Protein essentiality and predicted iMotifs.**

The fraction of yeast proteins that are products of essential genes [51] was calculated for the entire proteome, singlish-iMotif hubs (one or two iMotifs) and multi-interface hubs. iMotifs were assigned by applying the method to all yeast hubs in PIANA with *N* set to 20. The p-value of the difference between the whole data set and singlish- and multi-iMotif hubs (all-singlish and all-multi) and the singlish and multi-iMotif hubs (singlish-multi) was calculated using the Fisher's exact test for count data.

	Proteins tested for essentiality	Essential proteins	% essential	*P*-value
Entire proteome	6018	1116	19%	
All in PIANA	5034	1047	21%	
Singlish-iMotif hubs	90	27	30%	all-singlish: 0.04
Multi-iMotif hubs	507	262	52%	all-multi: 2.2×10^{-16} singlish-multi: 1.5×10^{-4}

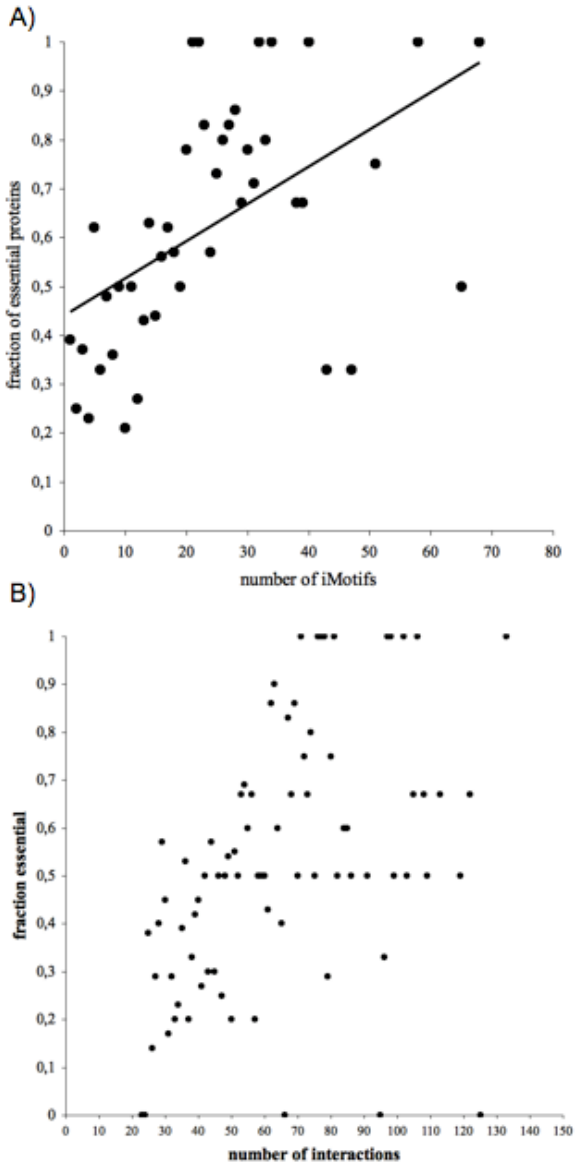


Figure 5. Essentiality study: proteins were binned according to their number of iMotifs (A) and to their number of interactions (B), and the fraction of essential proteins was calculated for each bin. Those bins with just one protein were not considered for calculating the correlations. A) Correlation between the number of iMotifs assigned to yeast hub proteins (≥ 20 interactions) in PIANA and the fraction of essential proteins (r_s is 0.61 and p-value is 1.6×10^{-5}). iMotifs were assigned to yeast hubs using an N threshold of 20. B) Correlation between the number of interactions of yeast hub proteins in Fig. 5A and the fraction of essential proteins (r_s is 0.51 and p-value is 1.1×10^{-6}).

Multi-iMotif hubs evolve slower than other proteins; singlish-iMotif hubs do not

A common measure of evolutionary rate is the dN/dS ratio (the ratio of non-synonymous to synonymous substitutions) [44]. Kim *et al.* found that multi-interface hubs have a lower evolutionary rate than the average protein in their data, but the same was not true for singlish-interface hubs. Our results are in agreement with their findings. Multi-iMotif hubs, in contrast to singlish-iMotif, evolve significantly slower than the average protein in our dataset (Table 2). However, the evolutionary rate difference between multi- (0.056) and singlish-iMotif hubs (0.062) was not found to be significant (p-value of 0.21).

Table 2. Protein evolutionary rate and predicted iMotifs.

The average evoluationary rate of yeast proteins [44] was calculated for the entire proteome, single-iMotif hubs, and multi-interface hubs. iMotifs were assigned by applying the method to all yeast proteins in PIANA with *N* set to 20. The p-value of the difference between the whole data set and singlish- and multi-iMotif hubs (all-singlish and all-multi) and the singlish- and multi-iMotif hubs (singlish-multi) was calculated using the Mann-Whitney U two-sided test.

	Entire proteome	All in PIANA	Singlish-iMotif hubs	*P*-value (all-singlish)	Multi-iMotif hubs	*P*-value (all-multi)	*P*-value (singlish-multi)
Evolutionary rate	0.077	0.074	0.062	0.12	0.056	8.1×10^{-11}	0.21

Discussion

We described, implemented, and evaluated a method that relies solely on binary protein interactions to identify interacting motifs (iMotifs) and their interactions. Our approach obtained high positive predictive value for identifying proteins with domains from the same SCOP family and predicting domain-domain interactions. We also analyzed hub proteins and their properties based on the number of iMotif assigned to them, obtaining similar findings to those in an independent approach that relies on protein structure information [14].

Recent estimates suggested that only one fifth of interaction types are known [38]. Therefore, current knowledge of protein structures is not sufficient to describe all protein interaction types. Our approach, in contrast to other previously described methods, accomplishes three different objectives: (i) it predicts the number of different iMotifs in a protein, (ii) classifies proteins by their predicted iMotifs, and (iii) predicts interactions between the iMotifs. The method can identify iMotifs independently of structural or sequence information; it can assign an iMotif to two structural domains or two iMotifs to a single domain. This property can be used to infer whether the interaction is mediated through multiple, single, or partial domains. The resolution at which iMotifs describe protein interfaces depends on the similarity metric used and the threshold applied by the method. On one hand, setting a high threshold on the number of common interaction partners (*N*) will assign few iMotifs to reduced sets of proteins (i.e. very specific and

restrictive iMotifs). On the other hand, using low *N* thresholds will assign the same iMotif to broad numbers of proteins (i.e. very unspecific and general iMotifs). We showed that the method works better for highly connected proteins and using high values for *N*. Moreover, our approach is not limited to finding predefined classes of protein components and thus allows us to predict new types of interacting motifs. On the one hand, an iMotif can be mapped to a predefined class (e.g., a SCOP domain or a PROSITE functional site) by examining the known classes assigned to proteins with that iMotif. On the other hand, iMotifs that remain unmapped are likely candidates for unknown classes. Such predictions may prove useful for target selection in structural genomics.

Relying solely on experimentally detected interactions affects the accuracy of our method. It has been shown that high-throughput experiments have limited reliability and that many of the detected interactions are probably not direct (i.e., they are carried out through a third protein) or do not even exist (i.e., false positives) [45]. However, we did not observe an improvement when solely using interactions from yeast two hybrid assays (Figure S2B), the high-throughput method that is best suited to discriminate between direct and indirect interactions. The question of restricting the method to use data from specific detection methods will have to be re-examined as more interaction data becomes available. One way of avoiding these limitations is to calculate similarity scores using families of proteins instead of absolute numbers of protein partners. This will prevent assigning the same iMotif to proteins that have many common partners but all of them belong to a single protein family. Removal of redundancy from the sets of partners indeed increases the percentage of identified protein pairs with a common domain family (Figure S6).

The iMotif assignments from our approach are similar to those obtained using an independent approach, which relies not only on known protein-protein interactions, but also on protein structure information [14]. In agreement with the results of Kim *et al.,* we observe different properties between hubs with multiple iMotifs (multi-iMotif) and hubs with one or two iMotifs (singlish-iMotif). In particular, we find that (i) multi-iMotif hubs are more likely than singlish-iMotif hubs to be essential for cell viability and (ii) multi-iMotif hubs, in contrast to singlish-iMotif hubs, evolve slower than the average protein. Furthermore, we have also observed a correlation between the number of iMotifs of a hub and its essentiality for cell survival. The properties observed for hubs with respect to their number of iMotifs may reflect the difference between proteins with multiple simultaneously possible interactions (multi-iMotif hubs are probably involved in permanent complexes) and proteins with multiple exclusive interactions (for singlish-iMotif hubs involved in transient

interactions). This is in agreement with the previous observation that interfaces of transient protein-protein interactions are less restricted in evolution than interfaces in permanent complexes [46].

Our results extend the findings and conclusions of Kim and co-workers [14] to proteins of unknown structure. Thus, inferring interacting motifs from protein interactions is likely to be helpful for providing biological insights about hubs for which no structural information is available.

Material and Methods

Protein interactions

Protein-protein interactions from DIP 2006.01.16 [47], MIPS 2006.01 [48], HPRD 2005.09.13 [49], BIND 2006.01 [50], and two recent high-throughput experiments [5,6] were integrated using PIANA version 1.2 [37], allowing us to work with a large set of 363,571 interactions between 42,040 proteins. PIANA represents protein interactions as a network where the nodes are proteins and the edges are interactions between the proteins. In such a network, a set of proteins linked to protein p_j (i.e., physically interacting with p_j) is named "partners of p_j". PIANA builds the protein interaction network by retrieving partners for a initial set of seed proteins. To avoid a positive bias in the method evaluation, interactions inferred from 3D structures were not used in this work.

Structural domains and Protein binding interfaces

Protein domain assignments and classification were obtained from the SCOP release 1.69 [22]. Here, domains are defined at the SCOP family level. Thus, domain-domain interactions refer to SCOP family interactions. The number of protein binding interfaces for hub proteins was obtained from the Structural Interaction Network 2.0 [14].

Essential Proteins and Evolutionary Rates

A list of ORFs essential for the survival of the yeast cell was obtained from the *Saccharomyces* Genome Deletion Project [51]. The evolutionary rates (dN/dS) of yeast proteins were taken from the adjusted values given by Wall et al. [44].

Assigning iMotifs to proteins and finding iMotif-iMotif interactions

The procedure is carried out in four steps:

1. Build the protein interaction network using the proteins of interest as seeds (see section "protein interactions" in Material and Methods).
2. Initialize a cluster interaction network (i.e., nodes are clusters that contain one or more proteins, and edges are interactions between clusters) by assigning each protein of the protein interaction network to a different cluster. In this initial cluster interaction network, each cluster (containing one protein p_j) interacts with those clusters that contain a partner of p_j in the protein interaction network.
3. Iteratively fuse the most similar clusters until the similarity score drops below a predefined threshold. The results presented in this work have been obtained using as similarity metric the number of common interaction partners (N). Therefore, the similarity between two clusters is their absolute number of common partners in the cluster interaction network. Other similarity metrics were considered, but none outperformed the use of N (Figure S1). When fusioning two clusters, the resulting cluster inherits the interactions that were common to both fused clusters. One protein can have multiple interfaces and therefore, in order to allow proteins to be in more than one cluster, clusters from the initial cluster interaction network (i.e., those that contain one single protein) are kept in the network even after being fused to another cluster.
4. Each cluster with more than one protein is labeled with a different interacting motif identifier (iMotif), and that iMotif is assigned to all proteins within that cluster. iMotif-iMotif interactions are derived from interactions in the cluster interaction network where both sides of the interaction have been labeled with an iMotif identifier.

For example (Figure 1), a proteome of six proteins (namely A, B, C, D, E, and F) forms a network of interactions that connects proteins A with B, C and D, and protein E with B, C, C, and F (step 1). Our method starts by creating a cluster interaction network from the network of protein interactions (i.e., 6 clusters with 7 interactions) (step 2). Next, the clusters that share the largest number of common interactions are fused (i.e., clusters 1 and 5, with 3 common interactions, are fused into a new cluster 6). This step is then repeated until the maximum similarity score between the clusters drops below a predefined threshold (i.e., 2 common interactions) (step 3). Thus, the iterative process will run for another iteration creating a new cluster (cluster 8) by fusing clusters 2, 3, and 4, which have two common interactions. Once the iterative process is finished, the method assigns iMotif

identifiers to clusters that contain more than one protein (i.e., cluster 7 becomes iMotif 1 containing proteins A and E, and cluster 8 becames iMotif 2 containing proteins B, C and D) (step 4). Moreover, iMotif-iMotif interactions are derived from the cluster interaction network (i.e. one interaction between iMotif 1 and iMotif 2).

Figure 2 illustrates iMotif assignments from a network perspective (Figure 2A) and from a structural perspective (Figure 2B). Moreover, the algorithm applied by the method is provided using pseudocode (Text S2).

Test set and evaluation procedure

We have evaluated the method on a test set created by selecting proteins (i) with at least 5 experimentally detected interactions, (ii) with at least 80% of their sequence covered by the domains defined in SCOP, and (iii) that did not introduce a redundancy bias in the evaluation (i.e., if any two sequences had a sequence identity greater than 30%, a BLAST e-value smaller than 10^{-5}, and the alignment had at least 30 residues, the shortest member of the pair was not selected). The final set contained 368 sequences (Table S4).

The SCOP family assignment was evaluated by considering as positive assignments those proteins found by the method to have a common iMotif with the query protein. Among these positives, we define as true positives those proteins that have a common SCOP family code with the query protein. Moreover, we define as false negatives the proteins that have the same SCOP family code as the query protein but were not found by the method to share an iMotif.

iMotif-iMotif interaction predictions were evaluated against interacting SCOP families obtained from the PDB. Two SCOP domains were considered to interact if they were co-crystallized and had at least two atoms within 5Å distance. Because we are interested in domain interactions at the protein-protein interaction level, we excluded intra-chain interactions from this set. Our method creates a list of putative domain-domain interactions for each predicted iMotif-iMotif interaction by assuming that all domains of the query protein with one iMotif interact with all domains of proteins with the other iMotif. In this context, we define as positive any iMotif-iMotif interaction where the query protein is involved. A positive is then considered a true prediction if at least one of its putative domain-domain interactions is observed in the PDB. Finally, false negatives are interactions observed in the PDB for SCOP families of the query protein that do not appear in any list of putative SCOP family interactions.

To avoid biases in the evaluation, only proteins from the test set (before removing redundancy) and their SCOP families were considered when counting positives and negatives. The positive predictive value is defined as the number of true positives over the total number of positives, and sensitivity is the number of true positives over the sum of true positives and false negatives. The positive predictive value and sensitivity were calculated with respect to the similarity score threshold used for stopping the clustering. We also define the applicability of the method as the percentage of proteins with at least one positive under a given threshold.

Statistical tests

All correlations were measured using the Spearman rank correlation coefficient (r_s). The assessment on whether two binomial samples of essentiality observations are significantly different was calculated using the Fisher's test. The assessment on whether two non-Gaussian samples of evolutionary rate observations come from the same distribution was calculated using the Mann-Whitney U two-sided test. Correlations and differences in the observations were considered significant for *p*-values lower than 0.05. All tests were performed using the implementation provided by *R* [52].

Supporting Information

Figure S1. The percentage of protein pairs having a domain of the same SCOP family is plotted as a function of their similarity scores (grouped in ranges of 10 units).

Figure S2. The percentage of protein pairs having a domain of the same SCOP family is plotted as a function of their similarity scores (grouped in ranges of 10 units), using the same parameters as in Fig. S1 but introducing new restrictions: A) proteins that have more than 70 interactions are ignored when performing the analysis; and B) only interactions from y2h are used.

Figure S3. Sample iMotif assignment (Text S1). A) Superposition of the prothrombin and the pancreatic trypsin inhibitor structures (PDB ids 1BTH and 2HPQ) shows an interaction through the SCOP family domain Eukaryotic proteases (in red). B) The structure of the anionic trypsin II interaction with the pancreatic trypsin inhibitor (PDB id 1BRB) also shows an interaction through the SCOP family domain Eukaryotic proteases (in red).

Figure S4. Performance of the method in predicting SCOP domain-domain interactions.

Figure S5. Spearman correlation coefficient between the number of interfaces and the number of iMotifs is plotted as a function of different *N* thresholds.

Figure S6. The percentage of protein pairs having a domain of the same SCOP family is plotted as a function of their similarity scores (grouped in ranges of 10 units), using the same parameters as in Fig. S1 but introducing a new restriction: redundancy was removed from the sets of partners to avoid artificial increase or decrease of the score caused by groups of homolog proteins.

Text S1. Example of iMotif assignment. To illustrate the method, we describe here a sample prediction for prothrombin (UniProt code THRB_HUMAN) (Figure S3).

Text S2. Algorithm for assigning iMotifs to proteins

Table S1. Number of protein pairs under each similarity score range for metrics described in Fig. S1.

Table S2. Complete list of iMotifs assignments for proteins in the test set.

Table S3. Complete list of number of iMotifs assigned to all hubs in PIANA.

Table S4. Proteins from the test set, using UniProt accession numbers

Data S1. iMotif assignments for all hub proteins in PIANA

Acknowledgements

We thank P.M. Kim for providing the data for Figure 4. We acknowledge all members of the Sali and SBI labs, especially Fred P. Davis and J. García, for helpful discussions and providing data. We also thank three anonymous reviewers for valuable comments.

Author contributions. RA conceived of the idea and performed research; AS, MMR and BO provided scientific guidance. RA, MMR and BO analyzed results; RA, AS, MMR and BO wrote the paper.

Funding. RA is supported by a grant from the Spanish Ministerio de Ciencia y Tecnología (MCyT, BIO2002-03609). The work has been supported by grants from the Spanish Ministerio de Educación y Ciencia (MEC, BIO02005-00533) and the ‘‘Programa Gaspar de Portolà (DURSI)’’; AS in particular was supported by NIH U54 RR022220 and PN2 EY016525, Sandler Family Supporting Foundation, IBM, Intel, Netapp, and Hewlett Packard.

Competing interests. The authors have declared no competing interests exist.

References

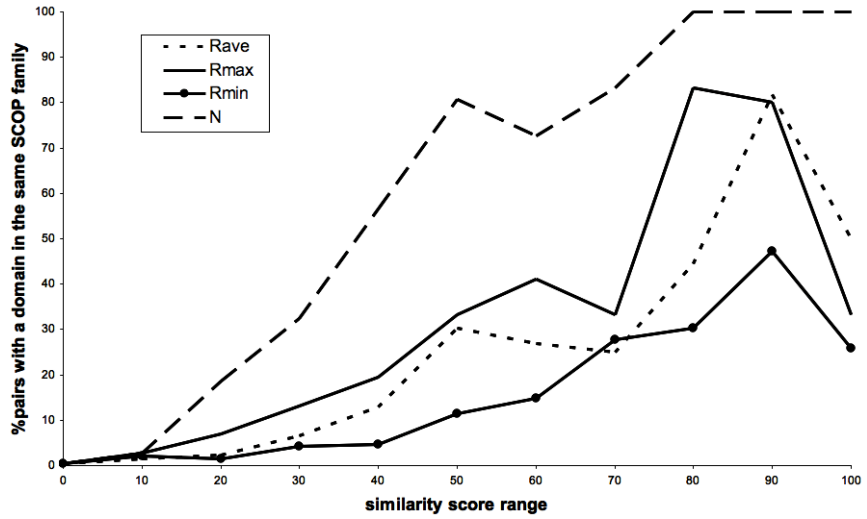
1. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147.
2. Pawson T, Gish GD, Nash P (2001) SH2 domains, interaction modules and cellular wiring. Trends Cell Biol 11: 504-511.
3. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7: 188-197.
4. Parrish JR, Gulyas KD, Finley RL, Jr. (2006) Yeast two-hybrid contributions to interactome mapping. Curr Opin Biotechnol 17: 387-393.
5. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.
6. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173-1178.
7. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24: 218-229.
8. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637-643.
9. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-636.
10. Liu J, Rost B (2003) Domains, motifs and clusters in the protein universe. Curr Opin Chem Biol 7: 5-11.
11. Kim WK, Henschel A, Winter C, Schroeder M (2006) The many faces of protein-protein interactions: A compendium of interface geometry. PLoS Comput Biol 2: e124.
12. Martin S, Roe D, Faulon JL (2005) Predicting protein-protein interactions using signature products. Bioinformatics 21: 218-226.
13. Espadaler J, Aragues R, Eswar N, Marti-Renom MA, Querol E, et al. (2005) Detecting remotely related proteins by their interactions and sequence similarity. Proc Natl Acad Sci U S A 102: 7151-7156.
14. Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314: 1938-1941.
15. Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res 33: D413-417.
16. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, et al. (2004) Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 342: 307-320.
17. Keskin O, Ma B, Nussinov R (2005) Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 345: 1281-1294.
18. Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 311: 681-692.

19. Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain-domain interactions from protein-protein interactions. Genome Res 12: 1540-1548.
20. Ng SK, Zhang Z, Tan SH (2003) Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19: 923-929.
21. Han DS, Kim HS, Jang WH, Lee SD, Suh JK (2004) PreSPI: a domain combination based prediction system for protein-protein interaction. Nucleic Acids Res 32: 6312-6320.
22. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
23. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.
24. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2005) InterPro, progress and status in 2005. Nucleic Acids Res 33: D201-205.
25. Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44: 336-343.
26. Li H, Li J (2005) Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. Bioinformatics 21: 314-324.
27. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins 58: 134-143.
28. Espadaler J, Romero-Isart O, Jackson RM, Oliva B (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. Bioinformatics 21: 3360-3368.
29. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41-42.
30. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430: 88-93.
31. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296: 750-752.
32. He X, Zhang J (2006) Why do hubs tend to be essential in protein networks? PLoS Genet 2: e88.
33. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, et al. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol 4: e317.
34. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, et al. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5: R6.
35. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, et al. (2003) From gene networks to gene function. Genome Res 13: 2568-2576.
36. Samanta MP, Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci U S A 100: 12579-12583.
37. Aragues R, Jaeggi D, Oliva B (2006) PIANA: protein interactions and network analysis. Bioinformatics 22: 1015-1017.
38. Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22: 1317-1321.
39. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399-403.

40. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. (2006) The PROSITE database. Nucleic Acids Res 34: D227-230.
41. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
42. Stelzl U, Wanker EE (2006) The value of high quality protein-protein interaction networks for systems biology. Curr Opin Chem Biol 10: 551-558.
43. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7: 120.
44. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483-5488.
45. Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein-protein interaction data? J Mol Biol 327: 919-923.
46. Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A 102: 10930-10935.
47. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449-451.
48. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, et al. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21: 832-834.
49. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32: D497-501.
50. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33: D418-424.
51. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.
52. R: A language and environment for statistical computing [http://www.r-project.org]. R Foundation for Statistical Computing, Vienna, Austria.

Selected supporting information

Figure S1. The percentage of protein pairs having a domain of the same SCOP family is plotted as a function of their similarity scores (grouped in ranges of 10 units).

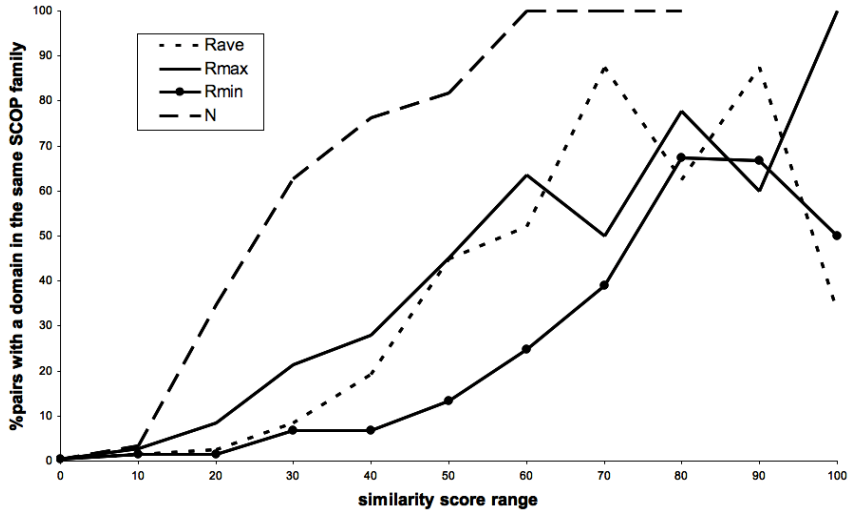


To measure the likelihood of two proteins p_i and p_j having a common interacting motif we defined four different similarity metrics: 1) N: the number of interaction partners that are common to p_i and p_j (long dashed line); 2) R_{max}: the ratio between N and the number of partners of the protein with more partners (bold line); 3) R_{min}: the ratio between N and the number of partners of the protein with fewer partners (circles); 4) R_{ave}: the average of metrics R_{max} and R_{min} (dotted line). For each score obtained using the similarity metrics described above, the percentage of protein pairs within that score range is plotted. For example, we observed that using N as the similarity metric, 73% of proteins with 50-60 common interaction partners shared a SCOP domain.

Table S1. Number of protein pairs under each similarity score range for metrics described in Fig. S1. In parenthesis, the number of pairs with at least one domain within the same SCOP family is indicated. We observe that metrics such as R_{min} outperform *N* at detecting a higher number of protein pairs with a domain within the same SCOP family, but this is done at the expense of being less precise.

Similarity Score range	R_{ave}	R_{max}	R_{min}	N
0	62296 (243)	62296 (243)	62296 (243)	62296 (243)
01-10	1678 (26)	3271 (92)	995 (21)	4487 (122)
11-20	1719 (41)	1053 (73)	1446 (22)	352 (66)
21-30	904 (60)	371 (49)	938 (39)	68 (22)
31-40	407 (53)	108 (21)	812 (37)	30 (17)
41-50	182 (55)	39 (13)	344 (39)	26 (21)
51-60	63 (17)	17 (7)	175 (26)	11 (8)
61-70	12 (3)	9 (3)	130 (36)	6 (5)
71-80	9 (4)	6 (5)	86 (26)	3 (3)
81-90	11 (9)	5 (4)	36 (17)	5 (5)
91-100	4 (2)	3 (1)	27 (7)	1 (1)

Figure S6: The percentage of protein pairs having a domain of the same SCOP family is plotted as a function of their similarity scores (grouped in ranges of 10 units), using the same parameters as in Fig. S1 but introducing a new restriction: redundancy was removed from the sets of partners to avoid artificial increase or decrease of the score caused by groups of homolog proteins. The procedure followed to remove redundancy was the same as the one used for creating the evaluation set. We observe a significant improve for all metrics with respect to Fig. S1.



CHAPTER IV

AN INTEGRATIVE APPROACH TO PREDICTING CANCER GENES

Cancer is a disease that causes thousands of deaths every year. In 2005, cancer became the leading cause of death in the United States for people under the age of 85 (American Cancer Society, Cancer statistics 2005 at http://www.cancer.org). However, recent developments in cancer research, prevention and treatment are showing the way towards a future in which cancer will be a marginal cause of death or life-quality loss.
In this chapter, we include an article submitted to BMC Bioinformatics in which we present the use of PIANA for the identification of proteins involved in cancer.
In this work, we can see the full potential of PIANA's capabilities in terms of integration: we have been able to simultaneously analyze data from protein-protein interaction databases, differential expression repositories, probabilities coming from a Naïve Bayes model, manually-curated lists, pathway information and protein annotations. The integration of multiple sources allowed us to obtain an accuracy significantly higher than methods that rely on a single source of data.

Articles included in this chapter:

Aragues R, Sander C, Oliva B. **An integrative approach to predicting cancer genes.** *Submitted to BMC Bioinformatics*

AN INTEGRATIVE APPROACH TO PREDICTING CANCER GENES

Ramon Aragues[1], Chris Sander[2], Baldo Oliva[1, §]

1.Structural Bioinformatics Lab. (GRIB). Universitat Pompeu Fabra-IMIM. Barcelona Research Park of Biomedicine (PRBB). 08003-Barcelona, Catalonia, Spain.

2.Computational Biology Center, Memorial Sloan-Kettering Cancer Center. 1275 York Avenue, Box 460, New York, NY 10021, USA

[§] Corresponding author

Email addresses: RA: ramon.aragues@upf.edu; CS: sanderc@mskcc.org; BO: boliva@imim.es

ABSTRACT

Background. Systematic approaches for identifying proteins involved in different types of cancer are needed. Experimental techniques such as microarrays are being used to characterize cancer, but validating their results can be a laborious task. Computational approaches are used to prioritize between candidate cancer genes, usually based on further analyzing experimental data.

Results. We implemented a systematic method using the PIANA software that predicts cancer genes by integrating data from multiple sources. Specifically, we produced lists of candidate cancer genes by relying on: (i) protein-protein interactions; (ii) differential expression data; and (iii) structural and functional properties of cancer genes. The integrative approach that combines multiple sources of data obtained positive predictive values ranging from 23% (on a list of 811 genes) to 73% (on a list of 22 genes), outperforming the use of any of the data sources alone. We analyze a list of 20 cancer gene predictions, finding that most of them have been recently linked to cancer in literature.

Conclusion. Our approach to identifying and prioritizing candidate cancer genes can be used to produce lists of genes likely to be involved in cancer. Our results suggest that differential expression studies yielding high numbers of candidate cancer genes can be filtered using protein interaction networks. We provide the complete list of human genes with the corresponding cancer gene prediction scores according to each type of data.

Background

Tumor development results from a progressive sequence of genetic and epigenetic alterations that promote the malignant transformation of the cell by disrupting key processes involved in normal growth control and tissue homeostasis [1]. Since complex biological networks control these processes, there are many genes that, mutated, can provide the cell with a specific aberrant capability. Alterations in three types of genes are responsible for tumorigenesis: oncogenes, tumor-suppressor genes, and stability genes [2]. Most oncogenes are involved in controlling the rate of cell growth, while tumor suppressor genes are usually negative regulators of growth or other functions that may affect invasive and metastatic potential, such as cell adhesion and regulation of protease activity. On the other hand, stability genes control the rate of DNA mutation, and their alteration can result in mutations in oncogenes or tumor suppressor genes, thus contributing to the development of cancer [3].

The completion of the human genome project and the development of high-throughput experimental techniques have enabled new approaches for studying cancer. For example, gene-expression profiling using microarrays has improved the classification of some tumor types [4, 5]. Moreover, data from large-scale screenings of protein-protein interactions has been used to identify interaction subnetworks activated in cancer [6]. Finally, genome scanning for gene copy-number alterations has detected many loci harboring candidate cancer genes [7]. Because of these advances, efforts to catalog all of the mutational events that contribute to human cancer can now be envisioned. For example, the Cancer Genome Atlas initiative (http://cancergenome.nih.gov) is resequencing a substantial fraction of human genes in order to elucidate the contribution of somatic mutations to cancer development and progression. Due to the complexity of these initiatives, methods to characterize and prioritize gene candidates likely to be involved in cancer must be developed [8-11].

Protein interaction networks have been shown to be a useful tool for better understanding the biology of the cell [12-14]. Moreover, the topology of the networks and the neighborhood of a given protein within the network have been used to functionally characterize proteins [15, 16]. It has also been observed that proteins related to a disease tend to have a high connectivity between them, specifically in inherited diseases [17, 18] and ataxia [19].

Gene expression profiling with DNA microarrays is a powerful approach for identifying cancer genes. Numerous studies have presented analyses of human cancer samples in which they identify gene expression signatures for different cancer types and subtypes [20-22]. In these experiments, genes are ranked according to their differential expression in the majority of cancer samples with respect to normal tissues, and genes above a predefined threshold are considered as candidate genes for the type of cancer being studied. Often, more in-depth analyses are performed to evaluate the involvement of candidate genes in the cancer, either by means of proteomics techniques [23], real-time polymerase chain reaction (qRT-PCR) [24], or literature search [25]. However, validating the results of microarray experiments can be a long and costly effort, due to the large number of candidate genes typically involved. Often, only a handful of genes of interest are selected for experimental validation, and hundreds of others are ignored. Moreover, due to limitations in DNA microarray technology, higher differential expressions do not necessarily reflect a greater likelihood of being related to cancer [26] and therefore, focusing only on the top candidate genes might not be the optimal procedure. Thus, there is a need for better techniques for selecting which differentially expressed genes will be analyzed in detail. Several procedures address the issue of cancer gene candidate selection [27] by further processing microarray data, either using more powerful statistics [28] or integrating multiple expression studies [29].

In order to improve the candidate gene selection process, several works have combined gene expression with other types of genomic data [30, 31]. One popular approach is gene set enrichment analysis, in which statistical tests are used to identify sets of dysregulated genes with a common biological function [32, 33]. Recently, Chinnaiyan and coworkers have combined the Molecular Concept Map and expression signatures to profile prostate cancer progression from benign epithelium to metastatic disease [34]. In the work of Rhodes *et al*. [6], instead of relying on predefined gene annotations, they applied a human interactome to genome-wide gene expression data in cancer for identifying a potential tumor suppressor gene in the integrin signaling pathway, and then demonstrated the utility of protein-protein interaction data for identifying interaction subnetworks activated in cancer. Finally, other approaches avoid the use of high throughput data by predicting cancer genes candidates based on their sequence, structure and functional properties [8, 35].

Here, we have implemented a systematic approach for identifying and comparing genes (and gene products) involved in cancer. Our method produces reduced lists of reliable candidate cancer genes by combining (i) a list of known cancer genes [10]; (ii) protein-protein interaction data [36]; (iii) expression information from multiple cancer studies [37];

and (iv) probabilities derived from structural and functional properties [35]. We begin by evaluating each method separately and comparing their results. Next, we present the integrative approach and evaluate its potential for predicting cancer genes. We provide candidate cancer genes obtained as a result of this work and assess them using public repositories of biological information and literature search. We conclude by discussing potential applications of our method.

Results

We were interested in assessing different methodologies for identifying cancer genes. Specifically, we tested the use of (i) protein interaction networks; (ii) microarray differential expression data; (iii) structural and functional properties of genes; and (iv) an integration of the three previous type of data. For the evaluation, we relied on a cancer gene list compiled from a variety of curated lists, cancer and sarcoma reviews, and Entrez Gene queries, followed by additional curation [10] (Material and Methods). We refer to genes in this list as the known cancer genes. Moreover, we use the term "cancer genes" to refer to genes and proteins involved in cancer.

Predicting cancer genes based on protein interaction partners

We assessed the use of protein interaction networks for predicting cancer genes. We hypothesized that proteins whose partners have been annotated as cancer genes are likely candidates for being cancer genes as well: if a mutated gene is perturbing a pathway related to cancer (e.g. growth control), mutations to interaction partners are also likely to perturb the same pathway. As corollary, proteins with many cancer genes interaction should be more likely to be involved in cancer than proteins with just one cancer gene partner. We used the PIANA (Protein Interactions And Network Analysis) tool [36] to build a cancer protein interaction network, using as seeds the gene products of the known cancer genes (Material and Methods). In this network, we define the cancer linker degree (CLD) of a protein as the number of cancer genes to which it is connected (Figure 1).

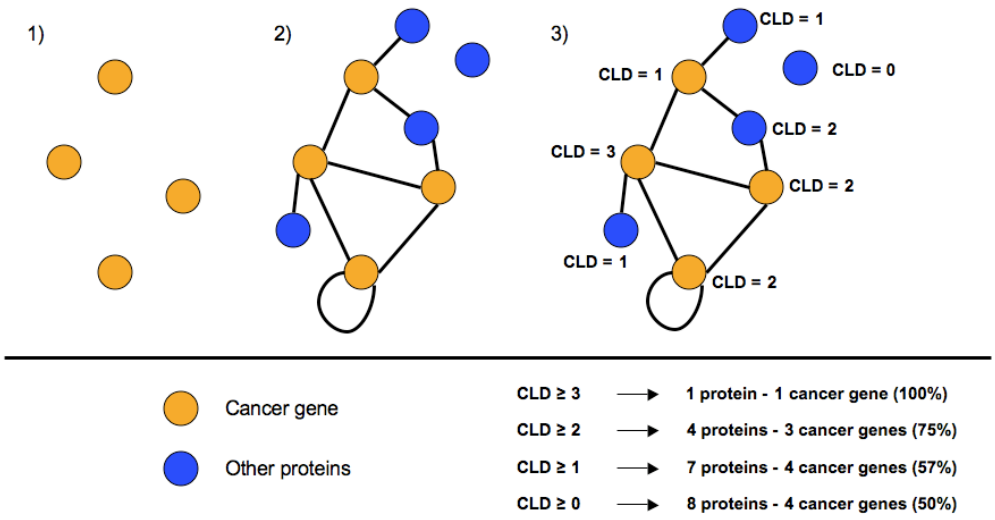


Figure 1. Calculating the Cancer Linker Degree (CLD) of a protein. The Cancer Linker Degree (CLD) of a protein is defined as the absolute number of partners of the protein that are known to be involved in cancer. In the example provided, we observe that proteins with high CLD are more likely to be cancer gene products that proteins with low CLD.

We examined the relationship between the CLD of a protein and its likelihood of being a known cancer gene. The fraction of cancer genes within proteins with CLD $\geq$ 10 is ~48%, compared to ~15% for proteins with CLD $\geq$ 1 or 10% for the average protein in our dataset (i.e. proteins with CLD $\geq$ 0). These results suggest that the Cancer Linker Degree of a protein is a good indicator of the probability of that protein being a cancer gene. We used the Cancer Linker Degree of a protein to predict cancer genes (Methods), obtaining a positive predictive value of ~54% at sensitivity of ~10% (Figure 2). Besides, similarly to previous studies [38], we observed that proteins with a large number of interaction partners (i.e., hubs) are more likely to be cancer genes than proteins with few interaction partners (see additional file 1). However, using the total number of interacting partners of a protein to predict cancer genes performed worse than using the cancer linker degree: for sensitivity of ~10%, the positive predictive value was ~35%.

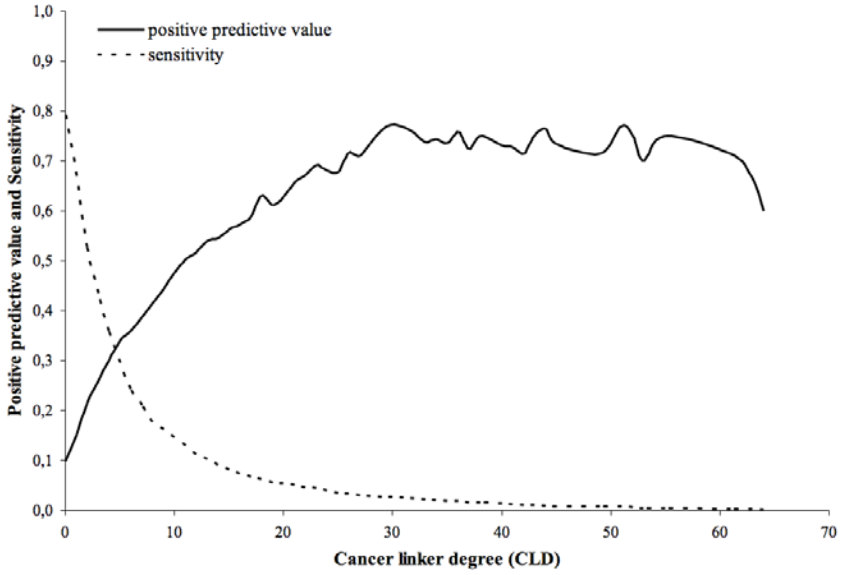


Figure 2. Positive predictive value and Sensitivity when predicting cancer genes based on the protein interaction partners in the cancer interaction network. The positive predictive value and sensitivity shown are for accumulative cancer linker degrees (CLD) (i.e. cancer linker degree 5 represents proteins with CLD ≥ 5). The average protein in the data set is represented by CLD 0.

Predicting cancer genes based on microarray data

We evaluated the use of differential expression data to predict cancer genes. We based our study on Oncomine [37] lists of over- and under-expressed genes in 24 differential expression studies, which we manually grouped in 12 different cancer types (see Material and Methods and additional file 2). The positive predictive value was between 9-16% for all cancer types, with sensitivity ranging from 84% (for genes over- or under-expressed in at least one cancer type) to 8% (for breast cancer) (Figure 3). We observed that genes appearing differentially expressed in multiple cancer types are more likely to be known cancer genes than those appearing differentially expressed in just one cancer type. For example, 20% of genes found differentially expressed in at least 5 cancer types are cancer genes, compared to 9% of genes found differentially expressed in at least one cancer type. These results confirm the need for post-processing in differential expression studies: microarrays detect many cancer genes, but they are usually mixed with many non-cancer genes.

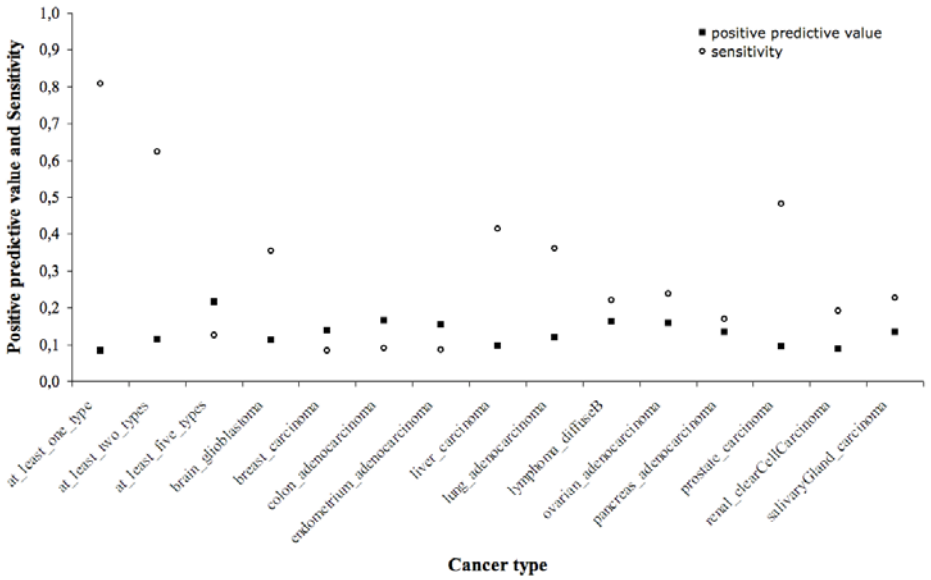


Figure 3. Positive predictive value and sensitivity when predicting cancer genes based on differential expression. The positive predictive value and sensitivity are shown for 12 cancer types and genes over- or under-expressed in at least 1, 2 and 5 cancer types.

Moreover, we studied the effect of looking at over- and under-expressed genes by their differential expression rank in a given experiment (i.e. their position in the list of over- or under- expressed genes ordered by their differential expression), as it is usually done in practice (Material and Methods). Among the 24 experiments tested, the positive predictive value when limiting the prediction to the 50 most differentially expressed genes outperformed the use of all differentially expressed genes in only 4 cases; in 10 cases it performed worse; and in the other 10 the positive predictive value was similar (see additional file 3). These results suggest that the number of cancer types in which a gene is observed differentially expressed is a better strategy for predicting cancer genes than using the differential expression rank of the gene.

Predicting cancer genes by structural, functional and evolutionary properties

Cancer genes have been shown to have common structural, functional and evolutionary properties [8, 35] and therefore, the properties of a gene can be used to estimate its probability of being a cancer gene [35]. We used the results from the work of López-Bigas and coworkers [35] to calculate the positive predictive value and sensitivity when predicting

cancer genes based on the structural, functional and evolutionary properties of genes (hereafter, we refer as SF-Probabilities to the probabilities assigned to genes in [35]). As shown on Figure 4, SF-Probabilities higher or equal to 0.90 yielded a positive predictive value of 21% at sensitivity of 13%, while for the average protein in the dataset (i.e. proteins with SF-Probability ≥ 0) the positive predictive value was 8% at sensitivity of 67%.

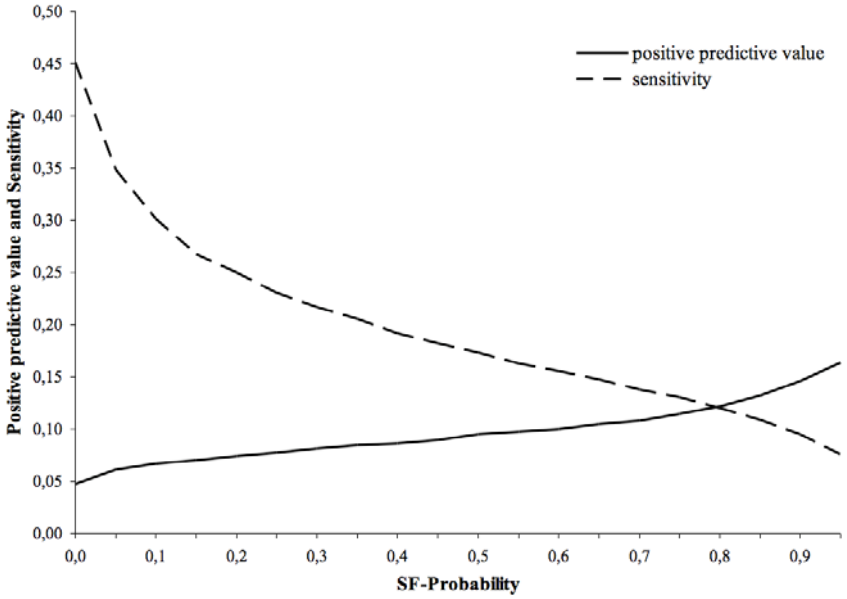


Figure 4. Positive predictive value and sensitivity are plotted as a function of the probability of being a cancer gene according to structural, functional and evolutionary properties (SF-Probability). The positive predictive value and sensitivity shown are for accumulative SF-Probabilities (i.e. SF-Probability 0.7 represents genes with SF-Probability ≥ 0.7). The average gene in the data set is represented by SF-Probability ≥ 0. SF-Probabilities were obtained from [35].

Relating the Cancer Linker Degree to differential expression and SF-Probability

Proteins with a high cancer linker degree tend to be differentially expressed in multiple cancer types

We were interested in examining the relationship between the cancer linker degree (CLD) of a protein and the number of cancer types in which its corresponding gene was differentially expressed. If proteins with high CLD tended to be differentially expressed in more cancer types than other proteins, that would suggest an involvement of high-CLD proteins in cancer. We observed that proteins with high CLD are more likely to be found differentially expressed in multiple cancer types than the average protein in the dataset (Figure 5):

proteins with CLD ≥ 1 appear differentially expressed in an average of 2.4 cancer types, compared to 4.4 cancer types for proteins with CLD ≥ 20. Furthermore, known cancer genes are found over- or under-expressed in an average of 2.8 cancer types.

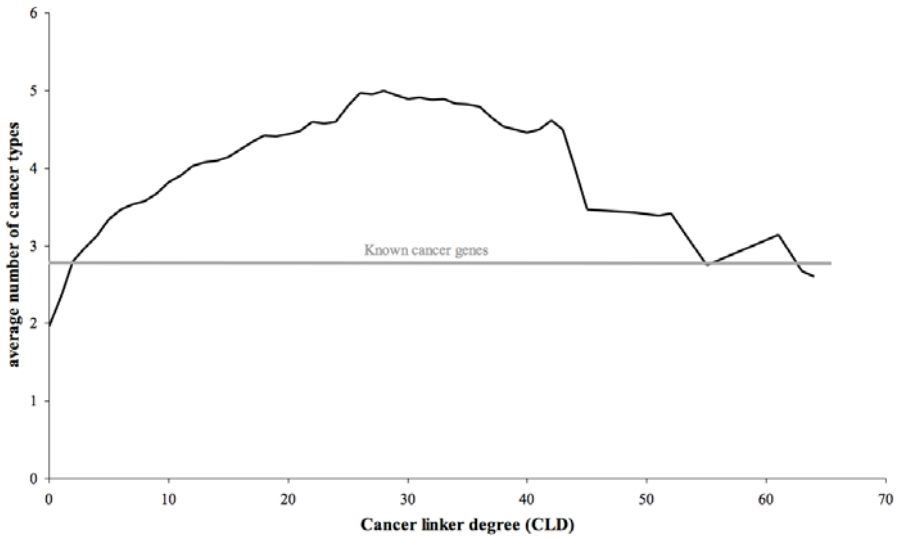


Figure 5. The average number of cancer types in which genes appear differentially expressed is plotted as a function of the cancer linker degree (CLD) of the gene products. The fractions of differentially expressed genes shown are for an accumulative CLD (i.e. CLD 5 represents proteins with CLD ≥ 5). The average protein in the dataset is represented by CLD 0. Known cancer genes appear differentially expressed in an average of 2.8 cancer types.

Proteins with a high cancer linker degree tend to have common functional, structural and evolutionary properties with cancer genes

We tested the correlation between the cancer linker degree (CLD) of proteins and their probabilities of being cancer genes according to their structural, functional and evolutionary properties (SF-Probabilities). We observed that proteins with high CLD tend to have higher SF-Probabilities than proteins with low CLD (Figure 6). For example, proteins with CLD ≥ 20 had an average SF-Probability of 0.51, compared to SF-Probability of 0.32 for proteins with CLD ≥ 1 or SF-Probability of 0.27 for proteins with CLD ≥ 0. Interestingly, proteins with CLD ≥ 3 had an average SF-Probability higher than that of known cancer genes. These results suggest that proteins highly connected to cancer genes in the cancer protein interaction network show structural, functional and evolutionary properties similar to cancer genes.

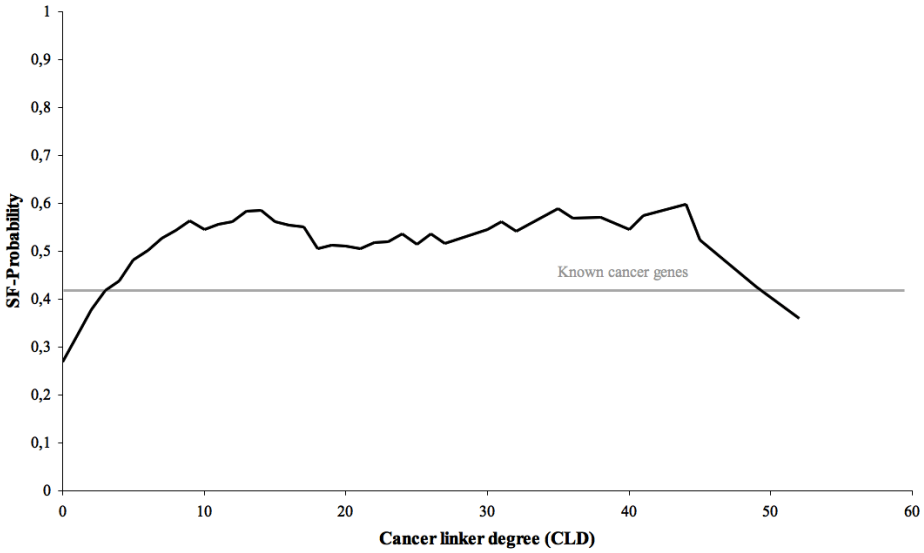


Figure 6. The probability of being a cancer gene according to structural, functional and evolutionary properties (SF-Probability) is plotted as a function of the cancer linker degree (CLD) of the gene products. The average SF-Probabilities shown are for an accumulative CLD (i.e. CLD 5 represents proteins with CLD ≥ 5). The average protein in the dataset is represented by CLD 0.

Predicting cancer genes by integrating multiple types of data

We evaluated the approach that predicts cancer genes using three different methodologies in conjunction: 1) the cancer linker degree (CLD) of proteins in the cancer protein interaction network; 2) the number of cancer types in which a gene appears differentially expressed with respect to normal tissue; and 3) the probability of being a cancer gene according to structural, functional and evolutionary properties (SF-Probability) [35]. The positive predictive values of this integrative approach ranges from 23% at sensitivity of 15% (for CLD ≥ 1, at least differentially expressed in one cancer type and SF-Probability ≥ 0.1) to 73% at sensitivity of 1% (for CLD ≥ 15, at least 5 cancer types and SF-Probability ≥ 0.0). Figure 7 shows the positive predictive value and sensitivity obtained when using multiple combinations of thresholds. We observed that the best results are obtained when combining a high CLD with the requirement of the gene being differentially expressed in at least 5 cancer types. Moreover, using high SF-Probability thresholds contributes towards reducing the number of false positives when applying lower CLD thresholds.

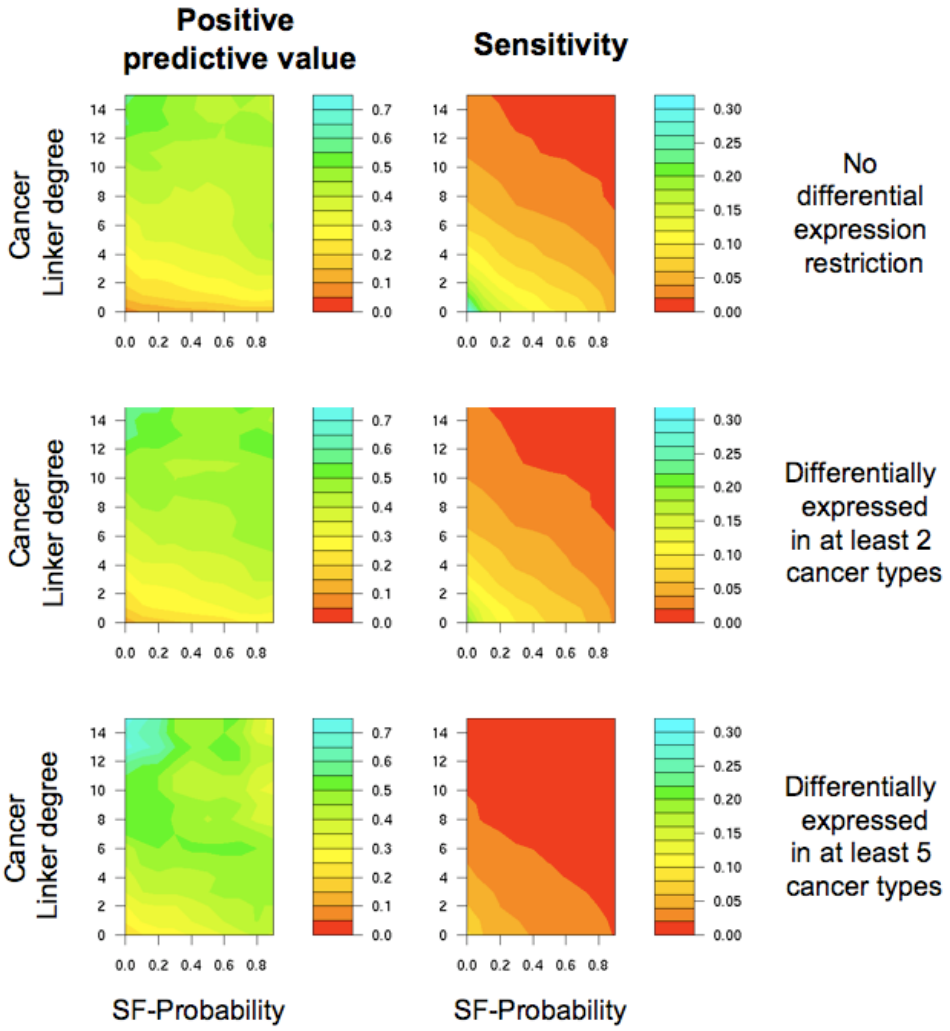


Figure 7. Contour maps for positive predictive value and sensitivity obtained when varying the thresholds applied by the integrative approach. In each image of Fig. 7, the x-axis is the SF-Probability threshold and the y-axis is the cancer linker degree (CLD) threshold. For a given restriction on the number of cancer types in which a gene must be differentially expressed in order to be considered a candidate, the positive predictive value and sensitivity are shown for each pair [CLD, SF-Probability]. Positive predictive values and sensitivities are shown using colored contour maps, from red (i.e. 0) to turquoise (i.e., 0.7 for positive predictive value and 0.3 for sensitivity). For example, imposing a gene to be differentially expressed in at least two cancer types, with a CLD of 6 and with an SF-Probability of 0.4, the positive predictive value is 0.4 for sensitivity of 0.05.

We also examined the contribution of each type of data by studying the positive predictive value obtained when segmenting the results by overlaps (Figure 8), observing that all criteria contribute towards obtaining reliable results. For example, the positive predictive value for each type of data use independently is 34% (for CLD≥ 5), 17% (differentially expressed in at least 4 cancer types) and 14% (for SF-Probability≥ 0.6); the overlap of the three criteria obtains a positive predictive value of 51%. These results show that the integrative approach can be used to produce reduced lists of reliable cancer gene candidates.

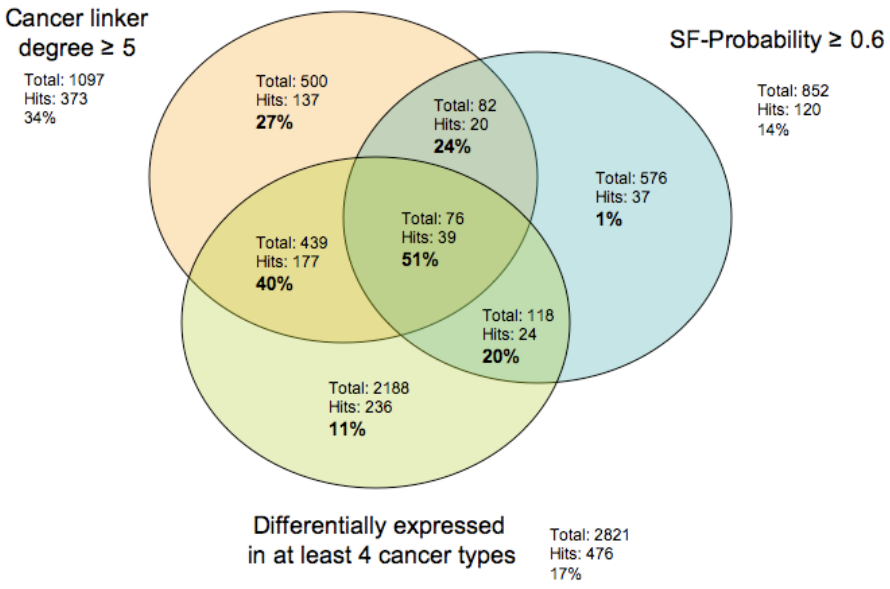


Figure 8. Positive predictive value calculated for diverse overlaps of cancer gene candidates. The criteria applied was the following: (i) cancer linker degree ≥ 5; (ii) differentially expressed in at least four cancer types; and (iii) SF-Probability ≥ 0.6. The Venn diagram shows the positive predictive value for cancer gene candidates predicted by applying the previous thresholds to the three types of data (i.e. 51%), and overlaps between each set of predictions.

Cancer gene candidates

The procedure followed to predict cancer gene candidates consists of four steps (Figure 9 and Methods): (i) using PIANA [36] to build the protein interaction network for the known cancer genes; (ii) mapping differentially expressed genes onto the network for each cancer type; (iii) mapping SF-Probabilities from [35] onto the network; (iv) producing an ordered list of candidates.

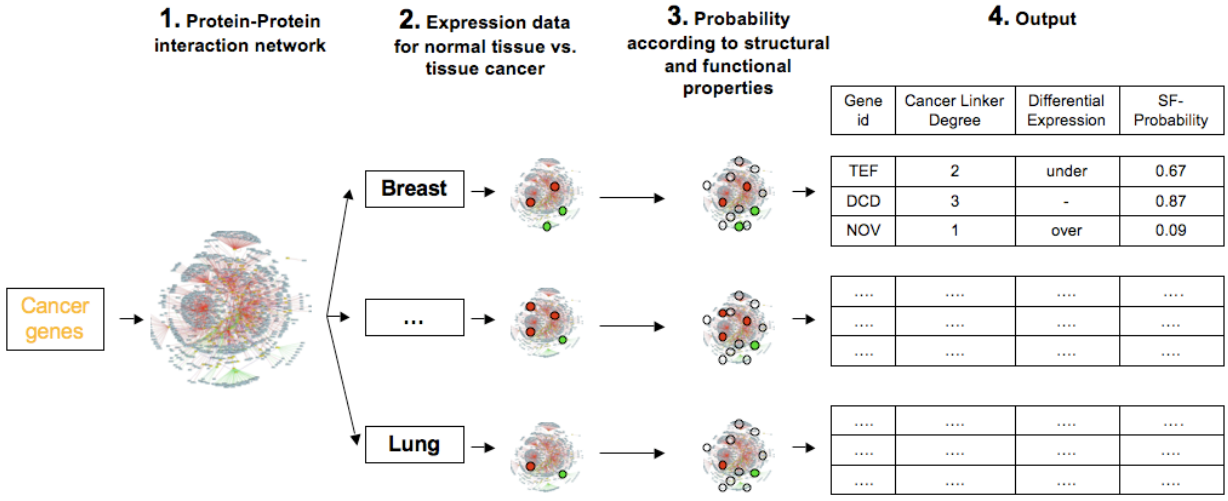


Figure 9. Procedure followed to predict cancer gene candidates. First, a cancer protein interaction network is built from the list of known cancer genes. Second, expression data from different cancer types is mapped onto the network. Third, probabilities of being a cancer gene based on structural, functional and evolutionary properties are retrieved for proteins in the network. Fourth, cancer genes are predicted based on the thresholds provided by the user for each type of data.

We have produced a reduced list of proteins likely to be involved in cancer (Table 1). Proteins in this list have a cancer linker degree equal or greater than 8, are differentially expressed in at least one cancer type and their SF-Probability is equal or greater than 0.70.

Table 1 can be used as a high-confidence resource for discovering new cancer genes. We also provide the complete list of human cancer gene candidates for which at least one type of data indicated a relationship to cancer (see additional file 4). This list comprises 11576 candidates, 1,040 of which where found by the three types of data (i.e. CLD $\geq$ 0, differentially expressed in at least one type of cancer and SF-Probability>0).

Table 1. Cancer gene candidates. The cancer gene candidates of this table where obtained by fixing the following thresholds: (i) cancer linker degree higher than 8; (ii) found differentially expressed in at least one cancer type; and (iii) probability based on structural, functional and evolutionary properties (SF-Probability) higher than 0.7.

Gene name	Cancer Linker degree	# of cancer types differentially expressed	SF-Probability
CDK9	11	6	0.97
GATA2	10	5	0.99
ATF2	17	6	0.94
CCNB1	13	3	0.73
CSNK2A2	22	4	0.89
PPARBP	14	5	0.99
CSK	19	5	0.90
KIN27	35	6	0.82
CUL1	12	3	0.85
DKFZP686I18166	11	6	0.99
STAT5B	20	6	0.99
MCM7	14	4	0.99
SURB7	14	4	0.74
MST1R	10	4	0.74
KHDRBS1	17	6	0.92
SYK	17	4	0.99
KDR	15	4	0.85
NME2	11	5	0.99
POLR2B	12	3	0.82
SRF	14	7	0.97

We analyzed (Table 2) cancer gene candidates from Table 1 based on literature search [39] and descriptions from the Cancer Gene Census [8], UniProt [40], Reactome [41] and the Gene Ontology (GO) [42]. This analysis suggests that our approach to identifying cancer genes is highly reliable: 60% of the proposed candidates have been directly related to cancer in experimental studies described in the literature, and an extra 25% participates in pathways known to be implicated in cancer. For example, the spleen tyrosine kinase (*syk*), predicted by the method to be a cancer gene, has been recently added (in a date subsequent to the creation of our list of known cancer genes) to the Sanger Cancer Gene Census [8]. *Syk*, with a cancer linker degree of 17, found differentially expressed in 4 types of cancer and with a SF-Probability of 0.99, is a positive effector of BCR-stimulated responses [43] and has been found to be involved in urinary bladder carcinoma [44] and primary liver cancer [45]. Besides, other candidate cancer genes have been related to cancer in the literature very recently (e.g. *mst1r*, involved in breast cancer [13]) or are known to be involved in pathways implicated in cancer (e.g. *srf* is a nuclear repressor of Smad3-mediated TGF-beta signaling [19], which induces apoptosis in numerous cell types). Finally, genes such as *surb7* and *kin27* were not found to be involved in cancer according to the literature, and future experimental studies should focus on evaluating their potential involvement in cancer. Literature references for each cancer gene candidate found to be involved in cancer are provided as additional file 5.

Table 2. Analysis of predicted cancer genes in Table 1. Column "related to cancer" indicates whether literature information [39], pathway membership and descriptions coming from UniProt and GenBank indicate a strong involvement in cancer (++), somehow related to cancer (+) or not related to cancer (-). Literature references for each gene found to be involved in cancer are provided as additional file 5.

Gene name	Description and Function/Pathway	Related to cancer
CDK9	Cell division protein kinase 9 Regulation of progression through cell cycle	++
GATA2	Endothelial transcription factor GATA-2 Transcriptional activator which regulates endothelin-1 gene expression	+
ATF2	Cyclic AMP-dependent transcription factor ATF-2 Transcriptional activator which binds to the CRE, present in many viral and cellular promoters.	+
CCNB1	G2/mitotic-specific cyclin-B1 Essential for the control of the cell cycle at the G2/M (mitosis) transition.	++
CSNK2A2	Casein kinase II subunit alpha Participates in Wnt signaling.	+
PPARBP	Peroxisome proliferator-activated receptor-binding protein Essential for embryogenesis. Plays a role in transcriptional coactivation	++
CSK	Tyrosine-protein kinase CSK Negative regulation of cell proliferation	++
KIN27	Protein kinase A-alpha ATP binding and protein serine/threonine kinase activity	-
CUL1	Cullin-1 Mediates the ubiquitination of proteins involved in cell cycle progression, signal transduction and transcription	++
DKFZP686I18166	Hypothetical protein ATP binding and protein kinase activity	-
STAT5B	Signal transducer and activator of transcription 5B Signal transduction and activation of transcription	++
MCM7	DNA replication licensing factor MCM7 Required for DNA replication and cell proliferation. Required for S-phase checkpoint activation upon UV-induced damage.	++
SURB7	Mediator of RNA polymerase II transcription subunit 21 Regulation of transcription.	-
MST1R	Macrophage-stimulating protein receptor [Precursor] Receptor for macrophage stimulating protein (MSP). Tyrosine-protein kinase activity.	++
KHDRBS1	KH domain-containing, RNA-binding, signal transduction-associated protein 1 Role in G2-M progression in the cell cycle.	++
SYK	Tyrosine-protein kinase SYK Positive effector of BCR-stimulated responses.	++
KDR	Kinase insert domain receptor Kinase activity and receptor activity.	++
NME2	Nucleoside diphosphate kinase B Major role in the synthesis of nucleoside triphosphates other than ATP.	++
POLR2B	DNA-directed RNA polymerase II 140 kDa polypeptide DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA.	+
SRF	Serum response factor SRF is a transcription factor that binds to the serum response element (SRE)	+

Discussion

We analyzed the use of three different criteria for predicting cancer gene candidates and concluded that: (i) the number of interaction partners of a protein that have been previously annotated as cancer gene (i.e. the cancer linker degree) is correlated with the likelihood of the protein to be involved in cancer; (ii) using differences in gene expression between normal tissue and cancer identifies many known cancer genes, but many non cancer genes as well; and (iii) probabilities based on structural, functional and evolutionary properties of known cancer genes (i.e. SF-Probabilities) are useful for filtering false positives from other cancer gene prediction methods. Moreover, we implemented and evaluated a method that integrates these criteria to produce reliable lists of cancer gene candidates, obtaining a positive predictive value of 71% when using very restrictive thresholds. Finally, we provided lists of cancer gene candidates and analyzed them using literature sources and information from public repositories, showing that our predictions are highly reliable.

Most methods used for predicting or prioritizing cancer gene candidates are biased towards genes that are well annotated and/or familiar to the researcher. This leaves unexplored many potential cancer gene candidates. However, high throughput genomic and proteomic work has now yielded relatively unbiased, although noisy, genome- and proteome-wide data sets. For example, expression studies produce large lists of over- and under-expressed genes, which are then prioritized by their differential expression rank, usually with help of a limited number of literature searches. Our integrative approach to finding cancer gene candidates can be used to obtain unbiased lists of cancer gene candidates by using the cancer linker degree of proteins to filter expression studies. We observed that the low positive predictive value obtained when using differential expression data alone (around 15% for most cancer types) shows a four-fold increase when combining it with protein-protein interaction data.

Separately, each of the criteria presented here for cancer gene candidate prediction has its limitations. First, methods based on protein interaction networks are limited by the fact that many cancers are the result of perturbations in the regulation of genes, which is not captured by protein-protein interaction data. Second, differential expression based methods have the drawback that many cancers are not the result of a differential expression in a particular gene, but rather on a mutation that prevents the gene product from performing its function. Finally, methods based on structural, functional and evolutionary properties are very dependent on existing functional annotations and their predictions are more stochastic than

based on biological observations. Therefore, integrative approaches for cancer gene prediction should also consider other types of information, such a gene regulatory networks [47] and gene copy-number alterations [7].

The methods presented here were evaluated by comparing their predictions with a list of oncogenes, tumor suppressors and stability genes [10]. This list attempts to be as comprehensive as possible, but two possible biases raise from its use: (i) not all methods cover the space of cancer genes to the same extent (e.g. the model used to calculate SF-Probabilities was trained on genes for which mutations have been causally implicated in cancer); and (ii) the method based on protein interaction networks heavily relies on the initial set of seed cancer genes and thus, genes isolated in the cancer network will never be pinpointed. An alternative approach to seeding our method with a list of known cancer genes is one where the seeds for building the protein interaction network are cancer-related proteins obtained with low-throughput experimental methods [48, 49]. This would remove the bias introduced by the input list of known cancer genes. Besides, we are mapping expression levels of mRNA onto a network of protein interactions. However, it is known that the mRNA expression levels do not always match the protein expression levels [50]. This will be solved by the use of techniques than measure the protein expression levels in specific cancers [51].

Conclusions

In conclusion, we showed that the integration of multiple sources of data is more reliable than the use of one single criteria to predicting cancer genes. Moreover, differential expression studies could benefit from the use of protein-protein interaction data to further validate their results. For example, combining the cancer linker degree of a protein with differential expression increased the fraction of known cancer genes within cancer gene candidates from 17% to 48%. Recently developed experimental techniques promise an increase in the amount of cancer data available, including regulatory events, tissue localization and protein expression. Systems capable of integrating all available sources of data will be fundamental to better understanding the mechanisms of cancer and other diseases.

Methods

Known cancer genes

We downloaded cancer genes from the Memorial Sloan Kettering computational biology website CancerGenes (http://cbio.mskcc.org/cancergenes/) as of January 2007. We collected a set of known cancer genes by querying the website for "oncogene", "tumor suppressor" and "stability". This list comprises 1256 cancer genes, in particular 385 oncogenes, 471 tumor suppressors and 494 stability genes (several genes belong to more than one category).

Protein Interaction Data

We used PIANA [36] to integrate human protein interaction data from DIP 2007.02.19 [52], MIPS 2007.04.03 [53], HPRD v6.01 [54], BIND 2007.04.03 [55], IntAct 2007.04.23 [56], BioGrid v2.026 [57] and MINT 2007.04.05 [58]. The integration of different sources of interactions into a single database allowed us to work with an extensive set of 110,457 human interactions between 36,900 different protein sequences.

PIANA represents the protein interaction data as a network where the nodes are proteins and the edges interactions between the proteins. In such a network, a set of proteins linked to protein p_j (ie, physically interacting with p_j) is named "partners of p_j". PIANA builds the network by retrieving partners for an initial set of seed proteins (i.e. the proteins of interest).

Expression data

We manually searched for gene expression studies between normal tissue and cancer in Oncomine [37], a cancer profiling database. We downloaded lists of over- and under-expressed genes from a total of 24 Oncomine studies, corresponding to 12 different cancer types (see additional file 2 for the list of experiments, the cancer type category assigned to them, and the total number of over- and under-expressed genes in each experiment). A gene was considered to have a significant differential expression if its Q value was lower than 0.05. Q values are obtained by correcting for multiple hypothesis testing the *p*-values calculated using Student's *t*-test for two-class differential expression analyses. A detailed description of the normalization process and statistical tests used in Oncomine can be found in [34, 37] .

Probabilities of being cancer-gene based on structural and functional properties

We used the probabilities of being a cancer gene calculated in [35] for all human genes. These probabilities were obtained using a Bayesian classification model that scored human genes for their likelihood of involvement in cancer according to structural, functional and evolutionary properties. Specifically, Lopez-Bigas and coworkers [35] relied on GO annotations [42] and sequence properties such as the extent of conservation, paralogy, and the lengths of proteins and genes. We refer to these estimated probabilities as SF-Probabilities. 12,194 human genes had an associated SF-Probability, 240 of which had been used to train the Bayesian model. 706 human genes had an SF-Probability higher than 0.95, and the SF-Probability was lower than 0.1 for 6288 human genes. Finally, 758 genes didn't have an associated protein sequence in PIANA and thus, were not used in this work.

Genes, proteins and identifiers

We used PIANA [36] to map expression data and SF-Probabilities onto the interaction network, in particular gene symbols coming from Oncomine expression studies and Ensembl identifiers coming from [35]. Throughout the text, we use the term 'cancer gene' to refer to any gene or protein involved in cancer.

Evaluating the use of protein interaction networks to identify cancer genes

The cancer protein interaction network was built using PIANA [36] by setting the list of known cancer genes as seeds (see "protein interaction data", Material and Methods). In this network, we define the cancer linker degree (CLD) of a protein as the number of cancer genes to which it is directly connected (Figure 1). The Cancer Linker Degree was calculated for each protein and proteins were binned by their CLDs. In this context, and given a CLD threshold of N, positives are proteins with CLD $\geq N$. True positives are known cancer genes among positives. False negatives are known cancer genes whose CLD is lower than N. The positive predictive value is defined as the ratio between true positives and positives. Sensitivity is the ratio between true positives and the sum of false negatives and true positives. Positive predictive values and sensitivities are shown in Figure 2 for CLD thresholds with at least 5 positives.

Evaluating the use of differential expression data to identify cancer genes

We calculated how many over- or under-expressed genes were known cancer genes for each cancer type described on additional file 2. Moreover, we tested how many genes

differentially expressed in at least 1-5 cancer types were known cancer genes. Besides, we evaluated the use of differential expression data by the rank of the cancer gene candidate in the original order from lists in Oncomine [37]. Genes from these lists were binned in incremental ranges of 50, and the number of known cancer genes was calculated for each bin incrementally. In this context, any differentially expressed gene is considered a positive. Among positives, we define as true positives those that are known cancer genes. False negatives are known cancer genes not found differentially expressed.

Evaluating the use of structural, functional and evolutionary properties to predict cancer genes

At any given SF-Probability threshold, positives are proteins with a SF-Probability above or equal to that threshold. Among positives, true positives are those that are known cancer genes. False negatives are known cancer genes not found above the SF-Probability threshold. Genes used for training the model in [35] were discarded for the evaluation.

Protein functions, pathways and literature

We manually analyzed cancer gene predictions by examining (i) the protein function and description as defined in UniProt [40]; (ii) the pathways in which the protein participated according to Reactome [41]; (iii) the molecular function and biological process as classified in the Gene Ontology (GO) [42]; and (iv) published articles retrieved using iHop [39].

Statistical tests

The correlation between the cancer linker degree of proteins and the fraction of known cancer genes among them was measured using the R [59] implementation of the Spearman rank correlation coefficient (*rho*).

ADDITIONAL FILES

Additional file 1 – Positive predictive value and Sensitivity when using the total number of partners of a protein to predict cancer genes. (see below)

Additional file 2 – Gene expression studies considered for this work. (not included in thesis; see article)

All 24 studies were downloaded from Oncomine (http://www.oncomine.org). The studies were manually grouped in 12 different cancer types. The number of over- and under-expressed genes is shown for each cancer type.

Additional file 3 – Fraction of cancer genes based on their rank in the over-and under-expressed lists of a microarray experiment. (not included in thesis; see article)

The specificity shown is for an accumulative rank (i.e. rank 50 represents ranks below or equal to 50). All results from all microarrays used for this work are presented here, divided in four figures for clarity's sake. We observed that, among the 24 experiments tested, the positive predictive value when limiting the prediction to the 50 most differentially expressed genes outperformed the use of all differentially expressed genes in only 4 cases; in 10 cases it performed worse; and in the other 10 the positive predictive value was similar.

Additional file 4 – Table with all cancer gene candidates. (not included in thesis; see article)

For each human gene with at least one data type indicated relationhip to cancer, this table shows the cancer linker degree (CLD), the number of cancer types in which it appears differentially expressed and its probability of being a cancer gene according to structural, functional and evolutionary properties (SF-Probability).

Additional file 5 – Sources of information for analysis of candidate cancer genes in Table 1 of the article. (see below)

AUTHORS' CONTRIBUTIONS

RA conceived of the idea and performed research; BO and CS provided scientific guidance. RA drafted the manuscript. BO helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We thank N. Lopez-Bigas for providing the SF-Probability data. We thank Carlos Rodriguez and all members of cbio at mskcc for helpful discussions and comments, especially Emek Demir, Robert Hoffmann, Doron Betel and Nikolaus Schultz. RA is supported by a grant from the Spanish Ministerio de Ciencia y Tecnología (MCyT, BIO2002-03609). The work has been supported by grants from the Spanish Ministerio de Educación y Ciencia (MEC, BIO02005-00533).

REFERENCES

1. Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**(1):57-70.
2. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control**. *Nat Med* 2004, **10**(8):789-799.
3. Bielas JH, Loeb KR, Rubin BP, True LD, Loeb LA: **Human cancers express a mutator phenotype**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(48):18238-18242.
4. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer**. *Nat Genet* 2004, **36**(10):1090-1098.
5. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM: **Concordance among gene-expression-based predictors for breast cancer**. *The New England journal of medicine* 2006, **355**(6):560-569.
6. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network**. *Nat Biotechnol* 2005, **23**(8):951-959.
7. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J *et al*: **Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation**. *Genome Res* 2003, **13**(10):2291-2305.
8. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nat Rev Cancer* 2004, **4**(3):177-183.
9. Hu P, Bader G, Wigle DA, Emili A: **Computational prediction of cancer-gene function**. *Nat Rev Cancer* 2007, **7**(1):23-34.
10. Higgins ME, Claremont M, Major JE, Sander C, Lash AE: **CancerGenes: a gene selection resource for cancer genome projects**. *Nucleic Acids Res* 2007, **35**(Database issue):D721-726.
11. Nguyen DX, Massague J: **Genetic determinants of cancer metastasis**. *Nature reviews* 2007, **8**(5):341-352.
12. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.
13. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**(6868):141-147.
14. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al*: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440**(7084):631-636.

15. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nat Biotechnol* 2006, **24**(4):427-433.
16. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast**. *Nat Biotechnol* 2000, **18**(12):1257-1261.
17. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B *et al*: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets**. *Nat Genet* 2006, **38**(3):285-293.
18. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N *et al*: **A human phenome-interactome network of protein complexes implicated in genetic disorders**. *Nat Biotechnol* 2007, **25**(3):309-316.
19. Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE *et al*: **A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration**. *Cell* 2006, **125**(4):801-814.
20. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer**. *Nature* 2001, **412**(6849):822-826.
21. Notterman DA, Alon U, Sierk AJ, Levine AJ: **Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays**. *Cancer Res* 2001, **61**(7):3124-3130.
22. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF, Jr., Hampton GM: **Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer**. *Cancer Res* 2001, **61**(16):5974-5978.
23. Cho WC: **Contribution of oncoproteomics to cancer biomarker discovery**. *Molecular cancer* 2007, **6**:25.
24. Kuo WP, Liu F, Trimarchi J, Punzo C, Lombardi M, Sarang J, Whipple ME, Maysuria M, Serikawa K, Lee SY *et al*: **A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies**. *Nature biotechnology* 2006, **24**(7):832-840.
25. Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery**. *Nature reviews* 2006, **7**(2):119-129.
26. Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA microarray measurements**. *Trends Genet* 2006, **22**(2):101-109.
27. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nat Rev Genet* 2006, **7**(1):55-65.
28. Mehta T, Tanik M, Allison DB: **Towards sound epistemological foundations of statistical methods for high-dimensional biology**. *Nat Genet* 2004, **36**(9):943-947.
29. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression**. *Proc Natl Acad Sci U S A* 2004, **101**(25):9309-9314.
30. Rhodes DR, Chinnaiyan AM: **Integrative analysis of the cancer transcriptome**. *Nat Genet* 2005, **37 Suppl**:S31-37.
31. Mathew JP, Taylor BS, Bader GD, Pyarajan S, Antoniotti M, Chinnaiyan AM, Sander C, Burakoff SJ, Mishra B: **From bytes to bedside: data integration and computational biology for translational cancer research**. *PLoS computational biology* 2007, **3**(2):e12.

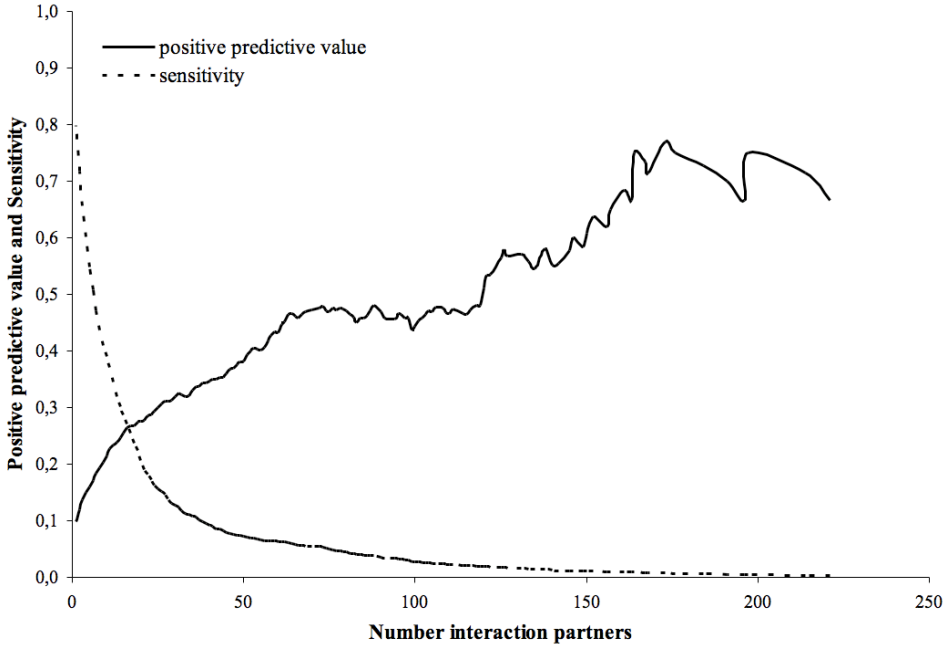
32. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**(3):267-273.
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
34. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ *et al*: **Integrative molecular concept modeling of prostate cancer progression**. *Nat Genet* 2007, **39**(1):41-51.
35. Furney SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N: **Structural and functional properties of genes involved in human cancer**. *BMC Genomics* 2006, **7**:3.
36. Aragues R, Jaeggi D, Oliva B: **PIANA: protein interactions and network analysis**. *Bioinformatics* 2006, **22**(8):1015-1017.
37. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **ONCOMINE: a cancer microarray database and integrated data-mining platform**. *Neoplasia* 2004, **6**(1):1-6.
38. Jonsson PF, Bates PA: **Global topological features of cancer proteins in the human interactome**. *Bioinformatics (Oxford, England)* 2006, **22**(18):2291-2297.
39. Hoffmann R, Valencia A: **A gene network for navigating the literature**. *Nature genetics* 2004, **36**(7):664.
40. **The Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2007, **35**(Database issue):D193-197.
41. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L *et al*: **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Res* 2005, **33**(Database issue):D428-432.
42. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D258-261.
43. Hong JJ, Yankee TM, Harrison ML, Geahlen RL: **Regulation of signaling in B cells through the phosphorylation of Syk on linker region tyrosines. A mechanism for negative signaling by the Lyn tyrosine kinase**. *The Journal of biological chemistry* 2002, **277**(35):31703-31714.
44. Kunze E, Wendt M, Schlott T: **Promoter hypermethylation of the 14-3-3 sigma, SYK and CAGE-1 genes is related to the various phenotypes of urinary bladder carcinomas and associated with progression of transitional cell carcinomas**. *International journal of molecular medicine* 2006, **18**(4):547-557.
45. Yuan Y, Wang J, Li J, Wang L, Li M, Yang Z, Zhang C, Dai JL: **Frequent epigenetic inactivation of spleen tyrosine kinase gene in human hepatocellular carcinoma**. *Clinical cancer research* 2006, **12**(22):6687-6695.
46. Kordes U, Hagel C: **Expression of SOX9 and SOX10 in central neuroepithelial tumor**. *Journal of neuro-oncology* 2006, **80**(2):151-155.
47. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C *et al*: **A genomic regulatory network for development**. *Science* 2002, **295**(5560):1669-1678.
48. Espana L, Martin B, Aragues R, Chiva C, Oliva B, Andreu D, Sierra A: **Bcl-x(L)-mediated changes in metabolic pathways of breast cancer cells: from survival in the blood stream to organ-specific metastasis**. *The American journal of pathology* 2005, **167**(4):1125-1137.

49. Mendez O, Martin B, Sanz R, Aragues R, Moreno V, Oliva B, Stresing V, Sierra A: **Underexpression of transcriptional regulators is common in metastatic breast cancer cells overexpressing Bcl-xL**. *Carcinogenesis* 2006, **27**(6):1169-1179.
50. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J *et al*: **Integrated genomic and proteomic analyses of gene expression in Mammalian cells**. *Mol Cell Proteomics* 2004, **3**(10):960-969.
51. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ *et al*: **Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression**. *Cancer Cell* 2005, **8**(5):393-406.
52. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.
53. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW *et al*: **The MIPS mammalian protein-protein interaction database**. *Bioinformatics* 2005, **21**(6):832-834.
54. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S *et al*: **Human protein reference database as a discovery resource for proteomics**. *Nucleic Acids Res* 2004, **32**(Database issue):D497-501.
55. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update**. *Nucleic Acids Res* 2005, **33**(Database issue):D418-424.
56. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R *et al*: **IntAct--open source resource for molecular interaction data**. *Nucleic Acids Res* 2007, **35**(Database issue):D561-565.
57. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Res* 2006, **34**(Database issue):D535-539.
58. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database**. *Nucleic Acids Res* 2007, **35**(Database issue):D572-574.
59. **R: A language and environment for statistical computing [http://www.r-project.orgl**. In. Edited by Team RDC: R Foundation for Statistical Computing, Vienna, Austria.

SELECTED ADDITIONAL FILES

Additional file 1 – Positive predictive value and Sensitivity when using the total number of partners of a protein to predict cancer genes.

We observed a clear distinction between proteins with many interaction partners and those with just a few partners. The positive predictive value and sensitivity shown are for accumulative numbers of partners (i.e. 'number of partners' 5 represents all proteins with 5 or more partners). Positive predictive value and sensitivity are shown for numbers of interaction partners with at least 5 positives.



Additional file 5 – Sources of information for analysis of candidate cancer genes in Table 1 of the article.

For each cancer gene candidate in Table 1 of the article, we reference one or more recent articles where the candidate has been linked to cancer. Information for all proteins was as well retrieved from UniProt [40] and from the literature using iHop [39].

Gene name	References
CDK9	[3]
GATA2	[4]
ATF2	[5]
CCNB1	[6]
CSNK2A2	[7]
PPARBP	[8]
CSK	[9]
KIN27	-
CUL1	[10]
DKFZP686I18166	-
STAT5B	[11]
MCM7	[12]
SURB7	-
MST1R	[13]
KHDRBS1	[14]
SYK	[15]
KDR	[16]
NME2	[17]
POLR2B	[18]
SRF	[19]

1. **The Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2007, **35**(Database issue):D193-197.
2. Hoffmann R, Valencia A: **A gene network for navigating the literature**. *Nature genetics* 2004, **36**(7):664.
3. Shan B, Zhuo Y, Chin D, Morris CA, Morris GF, Lasky JA: **Cyclin-dependent kinase 9 is required for tumor necrosis factor-alpha-stimulated matrix metalloproteinase-9 expression in human lung adenocarcinoma cells**. *The Journal of biological chemistry* 2005, **280**(2):1103-1111.
4. Li Z, Godinho FJ, Klusmann JH, Garriga-Canut M, Yu C, Orkin SH: **Developmental stage-selective effect of somatically mutated leukemogenic transcription factor GATA1**. *Nature genetics* 2005, **37**(6):613-619.
5. Liu H, Deng X, Shyu YJ, Li JJ, Taparowsky EJ, Hu CD: **Mutual regulation of c-Jun and ATF2 by transcriptional activation and subcellular localization**. *The EMBO journal* 2006, **25**(5):1058-1069.

6. Zhao M, Kim YT, Yoon BS, Kim SW, Kang MH, Kim SH, Kim JH, Kim JW, Park YW: **Expression profiling of cyclin B1 and D1 in cervical carcinoma**. *Experimental oncology* 2006, **28**(1):44-48.
7. Mazieres J, He B, You L, Xu Z, Jablons DM: **Wnt signaling in lung cancer**. *Cancer letters* 2005, **222**(1):1-10.
8. Zhu Y, Qi C, Jain S, Le Beau MM, Espinosa R, 3rd, Atkins GB, Lazar MA, Yeldandi AV, Rao MS, Reddy JK: **Amplification and overexpression of peroxisome proliferator-activated receptor binding protein (PBP/PPARBP) gene in breast cancer**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(19):10848-10853.
9. Humar B, Fukuzawa R, Blair V, Dunbier A, More H, Charlton A, Yang HK, Kim WH, Reeve AE, Martin I *et al*: **Destabilized adhesion in the gastric proliferative zone and c-Src kinase activation mark the development of early diffuse gastric cancer**. *Cancer research* 2007, **67**(6):2480-2489.
10. Nakayama KI, Nakayama K: **Ubiquitin ligases: cell-cycle control and cancer**. *Nature reviews* 2006, **6**(5):369-381.
11. Kazansky AV, Spencer DM, Greenberg NM: **Activation of signal transducer and activator of transcription 5 is required for progression of autochthonous prostate cancer: evidence from the transgenic adenocarcinoma of the mouse prostate system**. *Cancer research* 2003, **63**(24):8757-8762.
12. Kebebew E, Peng M, Reiff E, Duh QY, Clark OH, McMillan A: **Diagnostic and prognostic value of cell-cycle regulatory genes in malignant thyroid neoplasms**. *World journal of surgery* 2006, **30**(5):767-774.
13. Welm AL, Sneddon JB, Taylor C, Nuyten DS, van de Vijver MJ, Hasegawa BH, Bishop JM: **The macrophage-stimulating protein pathway promotes metastasis in a mouse model for breast cancer and predicts poor prognosis in humans**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(18):7570-7575.
14. Paronetto MP, Achsel T, Massiello A, Chalfant CE, Sette C: **The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x**. *The Journal of cell biology* 2007, **176**(7):929-939.
15. Yuan Y, Wang J, Li J, Wang L, Li M, Yang Z, Zhang C, Dai JL: **Frequent epigenetic inactivation of spleen tyrosine kinase gene in human hepatocellular carcinoma**. *Clinical cancer research* 2006, **12**(22):6687-6695.
16. Forsti A, Jin Q, Altieri A, Johansson R, Wagner K, Enquist K, Grzybowska E, Pamula J, Pekala W, Hallmans G *et al*: **Polymorphisms in the KDR and POSTN genes: association with breast cancer susceptibility and prognosis**. *Breast cancer research and treatment* 2007, **101**(1):83-93.
17. Ouatas T, Salerno M, Palmieri D, Steeg PS: **Basic and translational advances in cancer metastasis: Nm23**. *Journal of bioenergetics and biomembranes* 2003, **35**(1):73-79.
18. Michiels S, Danoy P, Dessen P, Bera A, Boulet T, Bouchardy C, Lathrop M, Sarasin A, Benhamou S: **Polymorphism discovery in 62 DNA repair genes and haplotype-associations with risks for lung, and head and neck cancers**. 2007.
19. Lee HJ, Yun CH, Lim SH, Kim BC, Baik KG, Kim JM, Kim WH, Kim SJ: **SRF is a nuclear repressor of Smad3-mediated TGF-beta signaling**. *Oncogene* 2007, **26**(2):173-185.

CHAPTER V

OTHER APPLICATIONS OF PROTEIN-PROTEIN INTERACTIONS

In this chapter, we include published articles in which the author of this thesis has been involved. In particular, we include the article from which the idea of creating PIANA was conceived (published in PNAS) and two articles in which PIANA was used in an experimental context: starting from data obtained in the wet lab, we made biological predictions that were later confirmed (or discarded) back in the bench.

Articles included in this chapter:

Espadaler J*, Aragues R*, Eswar N, Marti-Renom MA, Querol E, Aviles X, Sali A, Oliva B. **Detecting remotely related proteins by their interactions and sequence similarity.** Proc Natl Acad Sci U S A. 2005 May 17;102(20):7151-6

Mendez O, Martin B, Sanz R, Aragues R, Moreno V, Oliva B, Stresing V, Sierra A. **Underexpression of transcriptional regulators is common in metastatic breast cancer cells overexpressing Bcl-xL.** Carcinogenesis. 2006 Jun;27(6):1169-79

Espana L, Martin B, Aragues R, Chiva C, Oliva B, Andreu D, Sierra A. **Bcl-x(L)-mediated changes in metabolic pathways of breast cancer cells: from survival in the blood stream to organ-specific metastasis.** Am J Pathol. 2005 Oct;167(4):1125-37

* *Both authors contributed equally to this work.*

Detecting Remotely Related Proteins by Their Interactions and Sequence Similarity

Jordi Espadaler[1,2,#], Ramón Aragüés[1,#], Narayanan Eswar[3], Marc A. Marti-Renom[3], Enrique Querol[2], Francesc X. Avilés[2], Andrej Sali[3,*], and Baldomero Oliva[1,*].

1- Laboratori de Bioinformàtica Estructural (GRIB-IMIM). Departament de Ciències Experimentals i de la Salut. Universitat Pompeu Fabra. 08003-Barcelona, Catalonia, Spain.
2- Institut de Biotecnologia i Biomedicina and Departament de Bioquímica, Universitat Autònoma de Barcelona, 08193-Bellaterra (Barcelona), Spain.
3- Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA 94143-2240, USA.

Keywords: remote homology; fold assignment; family assignment; protein function annotation; protein-protein interactions.

Abbreviations: Structural Classification of Proteins (SCOP); Database of Interacting Proteins (DIP); Position Specific Scoring Matrix (PSSM).

# These authors contributed equally. * Corresponding Authors:

Baldomero Oliva. Laboratori de Bioinformàtica Estructural (GRIB-IMIM), Universitat Pompeu Fabra. C/ Doctor Aiguader, 80, Barcelona 08003, Catalonia, Spain. Tel: +34932240880; Fax: +34932240875; e-mail: boliva@imim.es

Andrej Sali. Mission Bay Genentech Hall, 600 16th Street, Suite N472D, San Francisco, CA 94143-2240, tel: +1 (415) 514 4227; fax: +1 (415) 514 4231, email sali@salilab.org, http::/salilab.org/

ABSTRACT

The function of an uncharacterized protein is usually inferred either from its homology to or interactions with characterized proteins. Here, we make use of both sequence similarity and protein interactions to identify relationships between remotely related protein sequences. We rely on the fact that homologous sequences share similar interactions, and therefore the set of interacting partners of a given protein's partners is enriched by its homologs. The approach was benchmarked by assigning the fold and functional family to test sequences of known structure. Specifically, we relied on 1,434 proteins with known folds, as defined in SCOP, and with known interacting partners, as defined in DIP. For this subset, the specificity of fold assignment was increased from 54% for PSI-BLAST to 75% for our approach, with a concomitant increase in sensitivity for a few percentage points. Similarly, the specificity of family assignment at the e-value threshold of 10^{-8} was increased from 70 to 87%. The proposed method will be a useful tool for large-scale automated discovery of remote relationships between protein sequences, given its unique reliance on sequence similarity and protein-protein interactions.

INTRODUCTION

Functional annotation of protein sequences by computation is essential in leveraging the impact of the genome-sequencing projects. To characterize the function of a protein sequence, it is often useful to identify its homologs and interacting proteins of known function. This task is facilitated by the classifications of protein domain families (1, 2), lists of protein-protein interactions (3, 4), and databases of protein structures (5-7). Protein domains are organized into *folds* (if sharing a similar structure), *superfamilies* (with evidence of homology in addition to structure similarity), and *families* (for homologs with similar function, sequence, and structure) (6). The vast majority of homologous sequences are expected to share the same fold.

The most sensitive algorithms for detecting homology between remotely related protein sequences rely on multiple sequence and protein structure information. The former group includes the sequence profile-based methods (8, 9) and Hidden Markov Models (10) that construct a multiple sequence alignment of the close homologs of the query, followed by scanning the corresponding profile against a database of sequences. The latter group includes sequence-structure threading methods that can sometimes reveal more distant relationships than purely sequence-based methods (11). Threading methods assign the fold by assessing the energy of coarse models corresponding to all possible ways of threading the sequence through each of the structures in a library of all known folds. Despite the increased coverage and accuracy of fold assignment using multiple sequence and structure information, two major problems remain for sequences related at less than ~25% sequence identity (12): (i) finding remote homologs that are undetectable by sequence similarity alone and (ii) identifying the functional family even when the fold can be detected (13, 14). Approximately ~60% of the known protein sequences have at least one domain with a reliable fold assignment, covering ~35% of the amino acid residues in the known protein sequences (15, 16).

Even when two sequences share little or no sequence similarity, their structures and functions may be similar (17, 18). Therefore, similarity in function may be indicative of a similar structure. An indicator of related functions is similar protein-protein interaction patterns. One such special case are the “interlogs” (*ie*, pairs of interacting proteins that interact identically in two species) (19). It has already been demonstrated that the information about the interacting partners can be used to predict the fold (7, 20-22) or function (14, 23-27) of a protein without considering its sequence. The utility of these approaches should grow with time, given the increasing amount of data about protein-

protein interactions (20, 24, 28), collected in databases such as BIND (3), MIPS (29), and DIP (4).

Our objective here is to demonstrate that the combined use of protein interactions and sequence similarity improves detection of remote similarity. We have implemented our method using the PSI-BLAST program, but any other method for detection of remote sequence similarity can be used. We begin by describing the approach. Next, we benchmark the method by relying on a set of non-redundant domains from SCOP that have known interacting partners defined in DIP. We conclude by discussing the implications of our results for protein structure modeling and functional annotation.

METHODS

A protein interaction network can be represented by a graph with nodes as proteins and edges as protein interactions. In such a graph, a set of proteins connected to protein X (*ie*, physically interacting with X) is named "partners of X". Moreover, we define successive levels of partnership: the set of partners of X is named "partners of X at level 1" and the set of partners of the partners of X at level 1 forms the set of partners at level 2, and so on. Given the commutative relation of the interactions (*ie*, if B is found in the set of partners of A, then A is found in the set of partners of B), protein X should be in the set of partners at level 2 of itself. In fact, protein X should occur in all sets of partners at even levels. Therefore, given that homologous proteins perform similar functions associated with similar interaction partners, the sets of partners of protein X at even levels contain more sequences homologous to protein X than a randomly selected set of sequences of the same size (Results). Furthermore, partners of protein X at levels 1 and 3 may also include some of its homologs because some proteins interact with their homologs or they evolved *via* a fusion of two genes of interacting ancestors (30). Here, we exploited these considerations in combination with sequence similarity to improve the assignment of a given protein sequence into the correct fold class and functional family.

We relied on three databases, the TrEMBL database of protein sequences (release 23.6 of April 2003) (31), the SCOP database of protein structure classification (version 1.65 from December 2003) (32), and the DIP database of experimentally identified protein interactions (release 20040113 of January 2004) (33).

The DIP database contains 16,903 protein sequences that are involved in 43,742 documented binary interactions. Fold, superfamily, and family domain codes of SCOP were assigned to a total of 4,324 proteins in DIP that could be matched by BLAST to a protein in

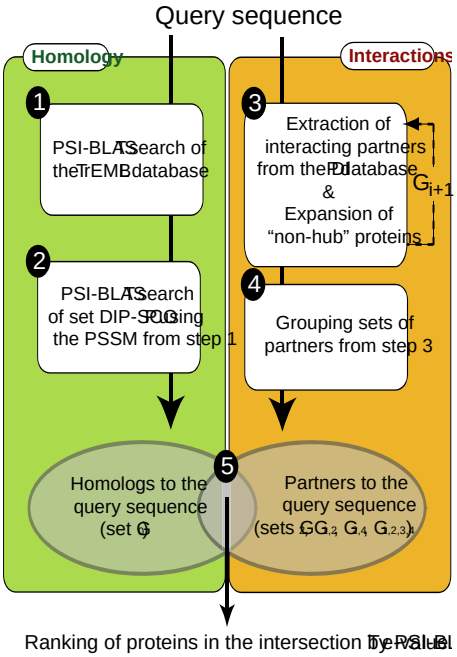
SCOP, covering one sixth of all proteins in DIP (*ie*, group DIP-SCOP). More precisely, one or more domain codes were assigned to a protein sequence in DIP when the alignment between the two sequences had an e-value smaller than 10^{-8} over at least 75% of the residues in the SCOP domain. A total of 4,743 binary interactions had SCOP codes for both proteins, while 14,813 interactions had the SCOP code for only one partner. This initial set of proteins was reduced to 1,434 query proteins to remove redundancies so that no two proteins in the set share more than 25% sequence identity after aligning them with the BLAST program.

Next, we added extrapolated links to the protein interaction network. Two proteins were linked by extrapolation if any members from their SCOP families interacted with each other. To enable benchmarking, the extrapolation was not performed for the query proteins in the benchmark. It was also not performed for "hub" proteins (34) that were defined here as proteins interacting with proteins in more than 10 different SCOP families. The hub proteins were excluded from extrapolation to minimize false positives. Thus, the list of reference links included both known interactions between pairs of domains as well as extrapolated links. Similarly, the term partner was expanded to include proteins connected *via* extrapolated links in addition to physical interactions.

The assignment of a fold or a family to a query sequence involves five steps (Figure 1). First, constructing a profile (PSSM) of the query sequence by searching for its homologs in the TrEMBL database (31) by PSI-BLAST(35) for a maximum of five iterations. Second, detecting query homologs in the DIP-SCOP group by PSI-BLAST using the query profile from step 1 (set G_0). Third, extracting partners of the query at levels 1, 2, 3, and 4 using the reference links. Fourth, grouping the sets of partners obtained in step 2 into four main groups, formed by the set of partners at level 2 (G_2), the union of the partners at levels 1 and 2 ($G_{1,2}$), the union of the partners at levels 2 and 4 ($G_{2,4}$), and the union of the four sets ($G_{1,2,3,4}$). Fifth, ranking the members of each of the groups in step 3 based on the e-value calculated in step 2. Additional combinations of partner levels are either redundant or complex, and are not reported in this study.

We tested family and fold assignment for different thresholds on the PSI-BLAST e-value with proteins in sets G_0, G_2, $G_{1,2}$, $G_{2,4,}$ and $G_{1,2,3,4}$. The number of positive assignments is defined as the number of sequences that align with the query sequence with an e-value smaller or equal to the threshold. Among these positives, we define the number of true positives as the number of sequences with the same SCOP code as the query sequence.

Figure 1. Flowchart for detection of remotely related proteins based on both sequence similarity and protein interactions.



First, for a query protein, a PSSM is built by five iterations of scanning the TrEMBL database by PSI-BLAST. Second, the PSSM is used in another PSI-BLAST run to obtain the e-values between the query and the proteins in DIP-SCOP. Third, the interaction partners of the query are extracted from DIP and may be expanded through SCOP family codes. This step is repeated, resulting in set G_i in iteration *i*. Fourth, partners at different levels are grouped as described in Methods. Fifth, proteins in the intersection are ranked by the PSI-BLAST e-value to the query, obtained in the second step.

RESULTS

Quantifying the enrichment afforded by protein interactions

Our method for detecting remote relationships by both sequence similarity and protein-protein interactions depends on the enrichment of the homologs among the set of partners of the query protein's partners (*ie*, the G_2 set). Therefore, we quantified this enrichment as follows. We first calculated the proportion of the correct fold assignments by dividing the sum of the correct fold assignments in each G_2 set by the sum of the G_2 set sizes (Table 1). Next, we compared this proportion against the corresponding proportion in the DIP-SCOP group. There is a significant enrichment of the proteins with the correct fold assignment in the G_2 set relative to DIP-SCOP. The same assessment was also performed for family assignment instead of fold assignment, revealing an even larger enrichment than that for fold assignment. Reflecting the homodimers, the corresponding statistics for the G_1 set also shows significant enrichment for homologs over a random selection from the DIP-SCOP set.

To quantify the statistical significance of enrichment in the G_2 set, we calculated the p-value with the Wilcoxon test(36). We compared for each query the enrichment in G_2 and in 1,000 random sets, with the same number of proteins as in group G_2, obtained from the DIP-

SCOP group. The corresponding p-value of 0.0064 quantifies the high statistical significance of enrichment in the G_2 set.

	G_1	G_2	DIP-SCOP
Fold assignment	0.137	0.041	0.018
Family assignment	0.107	0.018	0.003

Table 1. Enrichment for the correct folds and families. Proportions of the correct fold and family assignments in the G_1 and G_2 sets are compared with the proportion of the correct folds in the DIP-SCOP set.

Improved specificity of fold and family assignment

The specificity is defined as the number of true positives over the total number of positives. For an e-value cutoff below 1, our approach achieves ~75% specificity for group $G_{1,2}$ (Figure 2a). This relatively high specificity can be compared to the specificity of 54% for group G_0, obtained by PSI-BLAST alone; simultaneously, sensitivity is also improved for several percentage points (below). The improvement in specificity justifies the use of less significant e-value cutoffs in the filtered groups of sequences (G_2 and $G_{1,2}$) than with PSI-BLAST (G_0). The difference in performance between the traditional PSI-BLAST approach based on sequence matching alone and our approach, which also includes information about protein-protein interactions, increases as the e-value cutoff is raised.

The sets G_2 and $G_{1,2}$ were enriched for the correct family codes relative to the set G_0, demonstrating an improvement relative to searching by PSI-BLAST alone (not shown). The specificities obtained from groups G_2 and $G_{1,2}$ were ~80% for the e-value cutoff of 10^{-3}, while PSI-BLAST sequence search without consideration of interactions had the specificity of only ~60%.

Sensitivity of fold and family assignment

Our method cannot correctly assign a fold to a protein sequence when a stringent threshold on the E-value filters out correct predictions or when there are no experimental data about relevant protein interactions. To estimate the sensitivity, we defined as false negatives those undetected members of the same group that have the same domain fold as the query protein. Sensitivity for groups G_2 and $G_{1,2}$ is consistently better for several percentage points than for G_0 (Figure 2b).

Applicability of fold and family assignment

Our combined approach is not as general as the sequence comparison methods, which can be applied to all protein sequences. The reason is that the combined approach depends on the availability of protein interaction data. Therefore, to gauge the practical utility of the combined approach, we estimated its applicability to fold assignment for proteins in the sets G_2, $G_{1,2}$, $G_{2,4}$, and $G_{1,2,3,4}$ (Figure 2c).

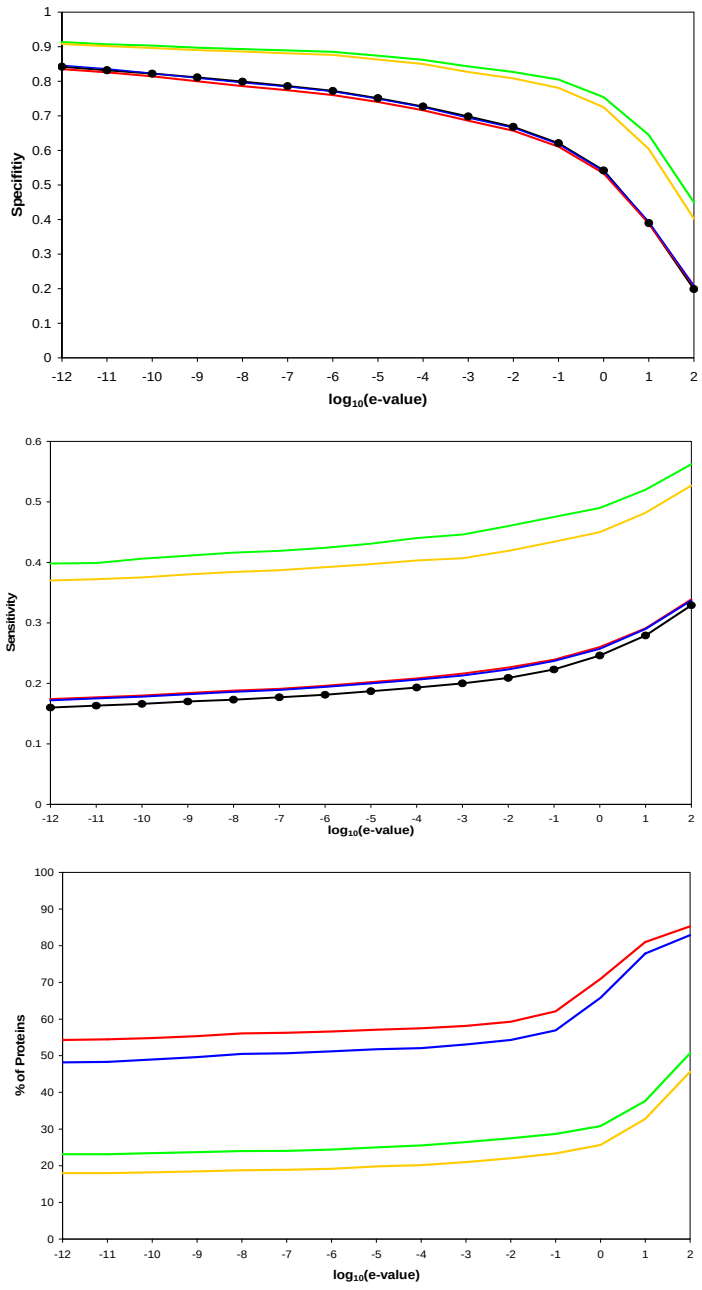


Figure 2. Specificity (a), sensitivity (b), and applicability (c) of fold assignment based on a combination of sequence similarity and protein interactions. The specificity, sensitivity and applicability are plotted as a function of the threshold on the PSI-BLAST e-value for groups G_2 (orange), $G_{1,2}$ (green), $G_{2,4}$ (blue), $G_{1,2,3,4}$ (red), and G_0 (black with circles).

Extrapolating interactions to increase the coverage

To assess the effect of the extrapolation of protein interactions (Methods), we compared the number of true positives at the fold level obtained with and without extrapolation, respectively. When used without extrapolation, our method is able to find only 286 true positives with 81% specificity at the PSI-BLAST e-value cutoff of 1, compared to 2,885 true positives and 75% specificity when using extrapolation. Thus, extrapolation yields a 10-fold increase in coverage with a relatively small loss in specificity. Even with extrapolation, however, only ~50% of the proteins in the DIP-SCOP group have a partner at level 2.

Example of fold and family assignment

To illustrate the ability of our approach to detect relationships between members of the same family in the absence of significant sequence similarity, we describe an example of the SwissProt sequences CNTF_HUMAN (ciliary neurotrophic factor) and ONCM_HUMAN (oncostatin M). CNTF_HUMAN is a survival factor for various neuronal cell types and ONCM_HUMAN is a growth regulator that inhibits the proliferation of a number of tumor cell lines. The two proteins share the same fold (4-helical cytokines) and family (long-chain cytokines). However, sequence similarity is very low (PSI-BLAST e-value is ~0.1; sequence identity is 16%).

According to DIP, both proteins interact with a member of the cytokine receptor family, LIFR_HUMAN (leukemia inhibitory factor receptor), as revealed by immuno-precipitation experiments (DIP entries 10988E and 10064E for the interactions of CNTF_HUMAN and ONCM_HUMAN, respectively). Moreover, the PDB structure 1I1R reveals a physical interaction between a member of the cytokine family (viral IL-6) and a member of the cytokine receptor family (human gp130). Thus, our method predicts with an e-value of 0.1 in group G_2 that CNTF_HUMAN has the same fold as ONCM_HUMAN (Figure 3a).

Example of fold assignment

To illustrate the ability of our approach to detect remote relationships at the fold level, we describe here an example of the Swiss-Prot sequences EF1G_YEAST and EF1B_YEAST. The C-terminal domains of these sequences adopt a ferredoxin-like fold. Nevertheless, EF1G_YEAST is an elongation factor 1γ of the eEF1-γ domain superfamily, while EF1B_YEAST is an elongation factor 1β of the eEF-1β-like superfamily. Both structures

share a core formed by a sheet of three β–strands and an external helix, and could be superimposed with an RMSD of 3.6 Å (Figure 3b). The e-value of the PSI-BLAST alignment between EF1G_YEAST and EF1B_YEAST is 0.036, obtained by querying the TrEMBL database with EF1G_YEAST for 5 iterations with default parameters.

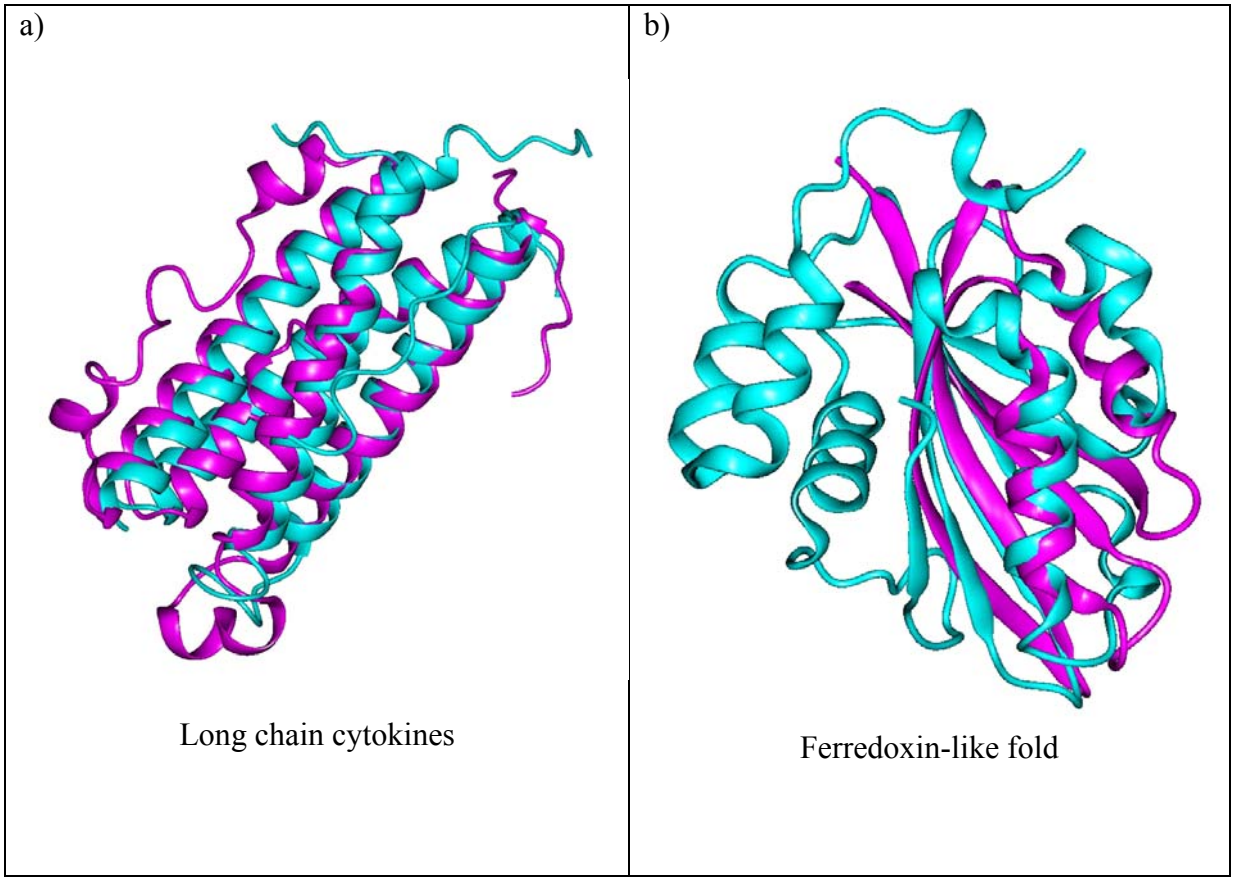


Figure 3. Illustration of fold assignment by the combined approach. (a) Structural superposition of the human ciliary neurotrophic factor (chain 1 of PDB code 1CNT, in cyan) and the human oncostatin *M* (chain A of PDB code 1EVS, in magenta). Structures were superposed with CE (53), with C_α RMSD of 1.7 Å and 15% sequence identity. (b) Structural superposition of two members of the ferredoxin-like fold, the C-terminal domains of human Elongation factor 1-γ (chain A of PDB code 1PBU, in cyan) and yeast Elongation factor 1-β (chain B of PDB code 1G7C, in magenta). The structures were superposed with CE, obtaining the C_α RMSD of 3.6 Å and 7.5% sequence identity.

The relationship between both sequences could be extracted through the interaction of EF1G_YEAST with an elongation factor 1α (EF1A_YEAST) in the G protein family, obtained from tandem affinity purification experiments (DIP entry 17026E, between nodes 6813N and 2250N). In addition, DIP contains an interaction between EF1B_YEAST and

TEM1_YEAST (a G protein) with the DIP entry code of 13895E (between nodes 6445N and 1691N), revealed by immuno-precipitation experiments. Therefore, EF1B_YEAST is a partner of EF1G_YEAST at level 2 (G_2). Table 2 shows the set of proteins found in G_0 of EF1G_YEAST with e-values between 10^{-3} and 1; there is only a single analog sharing the ferredoxin-like fold with EF1G_YEAST, and 2 false positives that do not appear in group $G_{1,2}$. In this case, our method is both sensitive and specific, because the correct fold of EF1G_YEAST appears in group $G_{1,2}$ without any false positives.

Table 2. Partial results from a search for homologs of EF1G_YEAST (folds 47615, 52832, and 54861; superfamilies 47616, 52833, and 89942) by PSI-BLAST. "Swiss-Prot", SwissProt codes of the sequences found in G_0 and aligned with the sequence of EF1G_YEAST with e-values between 10^{-3} and 0.1 (Methods). "e-value", the corresponding PSI-BLAST e-values. "Shares Fold" indicates whether or not the sequence shares a fold with EF1G_YEAST (ferredoxin-like fold). "SCOP Fold" and "SCOP Superfamily" indicate the SCOP fold and superfamily codes, respectively (multidomain proteins have multiple codes). "Appears in $G_{1,2}$" indicates whether or not a sequence is found in the $G_{1,2}$ set of EF1G_YEAST.

SwissProt	e-value	Shares Fold	SCOP Fold	SCOP Superfamily	Appears in $G_{1,2}$
SYEC_YEAST	0.027	no	52373	52374	no
EF1B_YEAST	0.036	yes	54861	54984	yes
SC14_YEAST	0.83	no	46928, 52086	46938, 52087	no

DISCUSSION

We described, implemented, and benchmarked a new method that uses information about both sequence similarity and protein-protein interactions to detect homology between remotely related protein sequences. The method was validated by a benchmark involving 1,434 query proteins of known structure (Figures 1 and 2) and illustrated by two examples (Figure 3). Although the method was benchmarked by using known protein structures, it is equally applicable to detection of remote relationships between protein sequences without known structures because it does not rely on protein structure information.

Generally, the function of uncharacterized proteins can be annotated in two fundamentally different ways (37). First, by establishing a sequence and/or structure similarity to another characterized protein; and second, by establishing a functional link to another characterized protein. The first group of methods includes sequence matching and threading (9, 11). The second group includes both experimental and computational methods, such as clustering by physical interactions (26), mRNA array expression profiles (38), analysis of gene fusion (30), phylogenetic profiles (39), and genomic association of genes (40). Our approach is unique in that it discovers homology by explicitly combining both sequence similarity and experimentally determined protein interactions. Therefore, it benefits from the databases of protein sequences, structures, and interactions. Another method infers fold and family membership from protein interactions (21), but not in combination with sequence similarity. While a few other sequence similarity-based methods, such as 3D-PSSM (22), also use functional information, this information is mined from scientific texts, not from lists of protein interactions.

The benchmark clearly suggests that protein interaction data increase the specificity and sensitivity of fold and family assignment (Figure 2). Consequently, our method allows the assignment of fold and family to a higher percentage of known protein sequences without loss of accuracy. For example, the specificity of fold assignment at the PSI-BLAST e-value cutoff of 1 was increased from 54% for PSI-BLAST to 75% when combining sequence similarity and protein-protein interactions, with a concomitant increase of sensitivity for several percentage points. Similarly, the specificity of family assignment at the e-value threshold of 10^{-8} was increased from 70 to 87%, also with a slight increase in sensitivity. Moreover, at the e-value cutoff of 1, more than 90% of the correct fold assignments share the same family as the query, while only 65% of the correct fold assignments with PSI-BLAST correspond to proteins with the same family code. This result was expected, given that our approach benefits from the conservation of interaction patterns usually related to the protein function and thus family classification.

The accuracy and coverage of our method are limited by false positives and negatives of sequence matching by PSI-BLAST (41, 42) as well as by false and missing interactions in DIP (43). To minimize sequence matching problems, additional methods, such as profile-profile searches (9), Hidden Markov Models (10), threading (44), and intermediate sequence search (41) can be used. As to the interactions, false positives rate and coverage can be

improved by probabilistic methods that rely on multiple sources of information about protein interactions (45) and by performing more experiments. Clearly, the coverage of the method will rise with the increase in the number of known protein-protein interactions that link query proteins to other proteins.

There are also intrinsic limitations of the method. For example, some of the proteins in the same SCOP family do not share the same interactions (46), resulting in false positives of our method. In addition, current interaction databases, including the DIP database, list protein-protein interactions, not domain-domain interactions. Therefore, the lack of distinction between a protein and a domain may also increase false positives when extrapolating links through the existence of common domains within proteins. This problem is reduced, but not eliminated, by not applying the extrapolation procedure to hub proteins.

The combined method is applicable only to protein sequences and their homologs for which protein interaction data are available, in contrast to sequence comparison alone, which is applicable to all protein sequences. This limitation is quantified by the following two examples. First, between ~20% and ~50% of the proteins in the DIP-SCOP group have a partner in the G_2 set (Figure 2c). Second, for specificity of 75%, sequence comparison by PSI-BLAST makes 30,302 pairs with correct fold assignments while our combined method finds 2,885 true positives of which 188 were not reported by PSI-BLAST. Two of these assignments are illustrated in Figure 3. We suggest that even the comparatively small coverage of the combined method is already useful in practice, given the two million known protein sequences that need to be related to each other; very few methods for characterization of proteins, experimental or computational, are applicable to most protein sequences and many proven methods are applicable to only a small fraction of all proteins. Moreover, the utility of our combined method is clearly increasing with the growth of the databases of known protein sequences and their interactions. We also expect that the idea of combining protein sequence comparison and protein interactions will enable additional future improvements in the matching of remotely related protein sequences.

There are several fold assignment methods, such as profile-profile matching, Hidden Markov Models, and threading, that are more sensitive than PSI-BLAST. We did not assess the performance of our approach against these methods because we focused on the relative utility of protein interactions when added to the consideration of sequence similarity.

However, we do suggest that our use of protein interactions will sometimes result in correct fold assignments when all other methods fail, especially when the most sensitive fold assignment methods are used instead of PSI-BLAST in our approach.

The proposed method is as applicable to establishing remote sequence-sequence matches as it is to fold assignment. However, we focused on fold assignment because of its importance in comparative protein structure modeling and structural genomics. The structural genomics initiative aims to experimentally determine carefully selected protein structures, such that most of the remaining sequences can be modeled with useful accuracy by comparative modeling (47). The number of experimentally determined structures for comparative modeling of most proteins based on at least 30% sequence identity to a known structure is estimated to be ~16,000 (48). A reduction of this number, while keeping the accuracy of the corresponding models constant, would reduce both the cost and time required by structural genomics to fulfill its aim (49-52). This reduction can be partly achieved by using more sensitive fold detection methods, such as the new method described here.

We plan to further develop our method to make it applicable to large-scale comparative protein structure modeling, and so increase the number of modeled proteins in MODBASE, our comprehensive database of comparative models for all known protein sequences that are detectably related to a known structure (15). The proposed method is expected to be a useful tool for large-scale automated discovery of remote protein similarities, given its unique reliance on sequence similarity and protein-protein interactions.

ACKNOWLEDGMENTS

We are grateful to D. Jaeggi, M.S. Madhusudhan, R.B. Russell, P. Aloy, C. von Mering, F. Davis, and H. Moss Sali for their comments. BO acknowledges a “Salvador de Madariaga” fellowship from the Spanish Ministerio de Educación, Cultura y Deporte (MECD). JE and RA acknowledge student fellowships of “Departament d’Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya” (DURSI) and Spanish Ministerio de Ciencia y Tecnología (MCyT), respectively. This work was supported by grants from Fundación Ramón Areces (BO), MCyT BIO2002-03609 (BO), “Programa Gaspar de Portolà del DURSI” (BO), NIH R01 GM54762 (AS), the Sandler Family Foundation (AS), Sun

Academic Equipment Grant EDUD-7824-020257-US (AS), an IBM SUR grant (AS), and an Intel computer hardware gift (AS).

REFERENCES

1. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004) *Nucleic Acids Res* **32 Database issue,** D138-41.
2. Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P. & Bork, P. (2004) *Nucleic Acids Res* **32 Database issue,** D142-4.
3. Bader, G. D., Betel, D. & Hogue, C. W. (2003) *Nucleic Acids Res* **31,** 248-250.
4. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004) *Nucleic Acids Res* **32 Database issue,** D449-51.
5. Pearl, F. M., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J. & Orengo, C. A. (2003) *Nucleic Acids Res.* **31,** 452-455.
6. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004) *Nucleic Acids Res* **32 Database issue,** D226-9.
7. Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. & Holm, L. (2001) *Nucleic Acids Res* **29,** 55-57.
8. Mittelman, D., Sadreyev, R. & Grishin, N. (2003) *Bioinformatics* **19,** 1531-1539.
9. Marti-Renom, M. A., Madhusudhan, M. S. & Sali, A. (2004) *Protein Sci* **13,** 1071-87.
10. Karchin, R., Cline, M., Mandel-Gutfreund, Y. & Karplus, K. (2003) *Proteins* **51,** 504-14.
11. Jones, D. T. (1997) *Curr.Opin.Struct.Biol.* **7,** 377-387.
12. Heger, A. & Holm, L. (2003) *J Mol Biol* **328,** 749-67.
13. Devos, D. & Valencia, A. (2001) *Trends Genet* **17,** 429-31.
14. Hegyi, H. & Gerstein, M. (2001) *Genome Res* **11,** 1632-40.
15. Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M. S., Davis, F. P., Stuart, A. C., Mirkovic, N., Rossi, A., Marti-Renom, M. A., Fiser, A., Webb, B., Greenblatt, D., Huang, C. C., Ferrin, T. E. & Sali, A. (2004) *Nucleic Acids Res* **32 Database issue,** D217-22.
16. Cherkasov, A. R. & Jones, S. J. (2004) *BMC Bioinformatics* **5,** 37.
17. Tian, W. & Skolnick, J. (2003) *J Mol Biol* **333,** 863-82.
18. Devos, D. & Valencia, A. (2000) *Proteins* **41,** 98-107.
19. Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S. & Vidal, M. (2001) *Genome Res* **11,** 2120-6.
20. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D. & Tyers, M. (2002) *Nature* **415,** 180-183.
21. Lappe, M., Park, J., Niggemann, O. & Holm, L. (2001) *Bioinformatics* **17 Suppl 1,** S149-56.
22. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000) *J Mol Biol* **299,** 499-520.

23. Letovsky, S. & Kasif, S. (2003) *Bioinformatics* **19 Suppl 1,** i197-204.
24. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002) *Nature* **415,** 141-147.
25. Samanta, M. P. & Liang, S. (2003) *Proc Natl Acad Sci U S A* **100,** 12579-83.
26. Fraser, A. G. & Marcotte, E. M. (2004) *Nat Genet* **36,** 559-64.
27. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003) *Nat Biotechnol* **21,** 697-700.
28. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J. M. (2000) *Nature* **403,** 623-627.
29. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. & Ruepp, A. (2004) *Nucleic Acids Res* **32 Database issue,** D41-4.
30. Marcotte, E., Pellegrini, M., Ho-Leung, Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999) *Science* **285,** 751-753.
31. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003) *Nucleic Acids Res* **31,** 365-70.
32. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. & Murzin, A. G. (2004) *Nucl. Acids. Res.* **32,** D226-D229.
33. Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J. & Eisenberg, D. (2004) *Nucleic Acids Res.* **32,** D449-451.
34. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001) *Nature* **411,** 41-42.
35. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res* **25,** 3389-3402.
36. Wilcoxon, F. (1945) *Biometrics* **1,** 80-83.
37. Sali, A. (1999) *Nature* **402,** 23, 25-23, 26.
38. Zhou, X., Kao, M. & Wong, W. (2002) *Proc Natl Acad Sci U S A.* **99,** 12783-12788.
39. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999) *Proc Natl Acad Sci U S A* **96,** 4285-8.
40. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998) *Trends Biochem Sci* **23,** 324-8.
41. John, B. & Sali, A. (2004) *Protein Sci* **13,** 54-62.
42. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998) *Proc Natl Acad Sci U S A* **95,** 6073-6078.
43. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002) *Mol Cell Proteomics* **1,** 349-56.
44. Zhang, C., Liu, S., Zhou, H. & Zhou, Y. (2004) *Protein Sci* **13,** 400-11.
45. Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. & Gerstein, M. (2002) *Trends Genet* **18,** 529-36.
46. Aloy, P. & Russell, R. B. (2002) *Trends Biochem.Sci* **27,** 633-638.
47. Baker, D. & Sali, A. (2001) *Science* **294,** 93-96.
48. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001) *Nat Struct Biol* **8,** 559-566.
49. Sali, A. (1998) *Nat.Struct.Biol.* **5,** 1029-1032.

50. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999) *Nat Genet* **23,** 151-157.
51. Burley, S. K. & Bonanno, J. B. (2003) *Methods Biochem Anal* **44,** 591-612.
52. Terwilliger, T. C., Park, M. S., Waldo, G. S., Berendzen, J., Hung, L. W., Kim, C. Y., Smith, C. V., Sacchettini, J. C., Bellinzoni, M., Bossi, R., De Rossi, E., Mattevi, A., Milano, A., Riccardi, G., Rizzi, M., Roberts, M. M., Coker, A. R., Fossati, G., Mascagni, P., Coates, A. R., Wood, S. P., Goulding, C. W., Apostol, M. I., Anderson, D. H., Gill, H. S., Eisenberg, D. S., Taneja, B., Mande, S., Pohl, E., Lamzin, V., Tucker, P., Wilmanns, M., Colovos, C., Meyer-Klaucke, W., Munro, A. W., McLean, K. J., Marshall, K. R., Leys, D., Yang, J. K., Yoon, H. J., Lee, B. I., Lee, M. G., Kwak, J. E., Han, B. W., Lee, J. Y., Baek, S. H., Suh, S. W., Komen, M. M., Arcus, V. L., Baker, E. N., Lott, J. S., Jacobs, W., Jr., Alber, T. & Rupp, B. (2003) *Tuberculosis (Edinb)* **83,** 223-49.
53. Shindyalov, I. N. & Bourne, P. E. (1998) *Protein Eng* **11,** 739-747.

España L, Martín B, Aragüés R, Chiva C, Oliva B, Andreu D, Sierra A.
Bcl-x(L)-mediated changes in metabolic pathways of breast cancer cells: from survival in the blood stream to organ-specific metastasis.
Am J Pathol. 2005 Oct;167(4):1125-37.

Méndez O, Martín B, Sanz R, Aragüés R, Moreno V, Oliva B, Stresing V, Sierra A.
Underexpression of transcriptional regulators is common in metastatic breast cancer cells overexpressing Bcl-xL.
Carcinogenesis. 2006 Jun;27(6):1169-79. Epub 2006 Feb 20.

CHAPTER VI

DISCUSSION

" The future has a way of arriving unannounced"

George Will

" The best thing about the future is that it only comes one day at a time"

Abraham Lincoln

"A conclusion is the place where you got tired of thinking"

Arthur Bloch

6.1 Discussion

We have presented the work developed during this thesis by means of published (and submitted) research articles. **Chapter I** places our work into context by (i) describing the biological processes behind proteins and their interactions; (ii) describing the experimental methods used for characterizing proteins and their interactions; (iii) providing an overview of biological databases; (iv) reviewing the state of the art in software platforms dedicated to protein interactions analysis, visualization and integration; and (v) introducing the use of bioinformatics methods for cancer diagnostic, prognosis and treatment. **Chapter II** describes PIANA, the software platform developed during this work. In the two articles included in Chapter II, PIANA is shown to be a useful tool for integrating data from multiple sources and using the data to provide biological insights about proteins and their interactions. Moreover, Chapter II provides a description of the protein interaction data available in the public domain and describes the properties of experimental and predicted protein interaction networks. **Chapter III** introduces and evaluates a method (implemented within PIANA) for identifying proteins that perform their interactions through the same interacting motif. Moreover, the results of this method are used to characterize protein hubs (i.e. proteins with many interaction partners). Specifically, Chapter III shows that some previously observed relationships between the number of interactions of a protein and genomic features such as essentiality and evolutionary rate are actually more reflective of the number of interacting motifs in the protein. **Chapter IV** illustrates the biomedical use of PIANA by predicting cancer gene candidates based on the integration of multiple types of biological data. The analysis of the results reveals that cancer genes predicted by multiple methods are more likely to be known cancer genes than those predicted by each method independently. Finally, **Chapter V** briefly introduces three research articles where PIANA was used in a wet lab environment and one research article published before PIANA was developed. In summary, this PhD thesis consisted in designing (and implementing) a software platform and using it in different research contexts.

One aspect to be highlighted for all the research articles included in this thesis is the importance of having access to as much information as possible. Most research results presented in the scientific literature are biased by the difficulties found to validate hypotheses on the full spectra of known data. The work performed during this thesis will

facilitate the use of all protein interaction data available, as demonstrated in the methods presented in Chapter III and IV. Moreover, the data integration approach presented here can be generalized to other types of biological data such as differential expression studies and gene regulatory networks. Having uniform and universal access to the plethora of data publicly available will be fundamental to the advance of molecular biology.

Each research article included in this thesis discusses the implications of the presented results. As in any scientific endeavor, finishing up a project is only the starting point for a new project. In this section, we briefly describe several directions in which this work could be continued.

6.1.1 Providing universal access to PIANA

PIANA is already being used in third-party laboratories across the world and publications referring and using our work are commencing to appear [187, 218-221]. However, PIANA is currently a tool for bioinformaticians and laboratories interested in using PIANA needs to perform a local installation of the platform and its database. In most experimental contexts, local installations are out of reach due to technical limitations and lack of expertise. In order to have a greater impact on the research community, new means of accessing PIANA should be available. For example, developing a web server capable of performing PIANA routines would be of great help for biologists interested in having access to all known interactions for their proteins of interest. A preliminary implementation of a web interface to PIANA has been developed [222], but its capabilities should be extended to allow multiple users and configurations.

One important aspect of open source software platforms is the involvement of the community within the development of the tools. PIANA features a highly modular architecture, and several developers have already contributed to improving it (e.g. [223]) However, in addition to participating in the development of PIANA, other researchers should be able to develop their own plug-ins to PIANA, in order to facilitate the implementation of new functionalities by external bioinformatics laboratories.

As a final remark, it is important to highlight that PIANA is already being used in wet lab environments to help guide decisions on which experiments should follow genomic and proteomic studies [224, 225]. We hope to extend the use of PIANA in experimental contexts by pursuing the goals described above. Bioinformaticians were not grown to help experimental biologists and producing data for bioinformaticians is not the ultimate goal of experimental biologists; both are there to contribute to the understanding of mechanisms

behind life, and it is the combination of all forces that will drive us to success. In particular, computational biologists should always have a double way of thinking: 1) helping in the advance of research in computational methods, models and data storage; and 2) facilitating to biologists the necessary means for easily accessing to bioinformatics tools. PIANA is a tiny, but correct, step towards both goals.

6.1.2 Representation and analysis of transient and permanent interactions

In the current PIANA data model, a protein interaction network is composed of nodes (proteins) and edges (interactions between proteins). However, many interactions taking place in the cell would be more faithfully represented by groups of proteins (complexes) that interact with another protein (or complex). Currently, we are working on labeling interactions in PIANA as 'permanent' or 'transient' (see section 1.2.1) based on information from external repositories. The next step will be to allow the creation of network nodes consisting of several proteins. In particular, PIANA should be able to represent an interaction where one protein complex interacts with another protein. Currently, the complex would appear as separate proteins in the network. A more adequate representation would be one where the complex proteins are presented as belonging to a single entity.

6.1.3 Biological Interactions And Network Analysis (BIANA)

PIANA currently works at the protein-protein interaction level. However, many molecular interactions involve other biological entities such as DNA (e.g. transcription factors interact with DNA to regulate gene expression) or RNA (e.g. complexes such as the ribosome include protein-protein interactions, RNA-RNA interactions and protein-RNA interactions). Moreover, many biological processes are influenced by factors other than direct physical contacts (e.g. phosphorylation). Finally, gene regulatory networks are fundamental to understanding the biology of the cell. PIANA should be able to handle these and other interactions taking place in the cell. This goal is currently being pursued and a preliminary redesign of the database behind PIANA has been performed to facilitate the transition from Protein Interactions And Network Analysis (PIANA) to Biological Interactions And Network Analysis (BIANA).

One important limitation of all software platforms designed to work with protein-protein interaction data is that they see the network as a static entity. However, biological processes

are dynamic and depend on factors such as presence/absence of inhibitors and protein concentrations (e.g., many transient interactions won't occur unless a minimum level of protein concentration is reached). Strategies to address these concepts should be implemented within PIANA. A first step towards this goal has been taken by allowing the user to restrict the view of the network to proteins within the same cellular compartment, or to those found co-expressed by gene expression studies. In the near future, we plan to move one step more towards dynamic networks by integrating into a single network protein interaction data with co-regulation information.

All of the improvements highlighted here will imply soft modifications to the PIANA software architecture, a slight redesign of the database behind PIANA, the development of new parsers and the implementation of classes and methods that deal with new biomolecules and interactions. However, none of these modifications will imply any substantial changes to the fundaments of PIANA.

6.1.4 Interaction Confidence Score

PIANA is an important contribution towards creating tools that facilitate the correct storage, analysis and use of the interaction data available. However, most interaction data within PIANA comes from high-throughput the reliability of high-throughput methods (HT methods) for detecting protein-protein interactions has been questioned [43-45]. For example, a number of false positives in yeast two hybrid are plausible interactions (i.e. the two proteins do interact when one is facing the other) that do not take place in vivo (e.g. they are never coexpressed in the cell), while many interactions reported with the Tandem Affinity Purification method are in fact a mere indication of two proteins participating in the same complex (see section 1.3.7 and [41, 43]). Different methodologies have been proposed to tackle the reliability of protein interaction data (see section 1.3.7), such as the "interaction generality" [61] and the IRAP* [62] methods.

Most protein interaction repositories do not provide interaction confidence scores, and therefore, PIANA considers all interactions as being equally reliable. We are currently working on the implementation of a robust reliability score for interactions within PIANA, in order to allow outputting results at different levels of confidence.

6.1.5 Visualization of protein interaction networks

Visualization is a vital aid in integrating and interpreting molecular interaction networks [142]. Currently, PIANA relies on third-party software for visually interpreting protein interaction networks. For example, PIANA produces outputs compatible with Cytoscape [184], a software environment for integrated models of biomolecular interaction networks. There are no plans to implement visualization capabilities within PIANA, but a more direct communication between PIANA outputs and visualization software should be addressed when PIANA becomes a web service (see setion 6.1.1). For example, implementing a plug-for Cytoscape that allows users to use PIANA would be a convenient way of benefiting from both Cytoscape visualization tools and PIANA integration and analysis capabilities.

6.1.6 Integrating sequence information into the method for delineating interacting motifs

The method described in Chapter III delineates protein interacting motifs by relying on the observation that proteins with common interaction partners tend to interact with those partners through the same interacting motif. In this method, the protein interacting motifs (referred as iMotifs) are inferred from protein interaction data alone; no sequence or structure information is needed. However, a preliminary study on the combination of sequence and the inferred iMotifs showed promising results, as many iMotif assignments could be confirmed or corrected by aligning the proteins within the same iMotif category. Therefore, future implementations of the method in Chapter III will have to take into account the sequence of the proteins at the time of performing the clustering. Briefly, the method will consist of the following steps: 1) build the protein interaction network; 2) create the initial cluster interaction network by assigning one protein to each cluster; 3) iteratively fuse the most similar clusters until the similarity score drops below a predefined threshold. In this step, the similarity metric between two clusters will be calculated based on (i) the common interaction partners of the clusters and (ii) the multiple alignment obtained from all proteins within both clusters; 4) assign one iMotif identifier to each cluster with more than one protein and derive iMotif-iMotif interactions. In this step, two clusters with high overlap in terms of protein members will be fused if a common sequence pattern can be found between the two sets of proteins.

Structure-based methods are the best placed for correctly identifying and classifying protein interfaces [226, 227]. However, these methods are limited by the relatively low number of

known 3D structures for proteins and their complexes, which results in a low coverage of the space of proteins and their interactions. Therefore, the question of combining iMotif assignments with structure-based approaches for the identification of binding sites should also be considered. For example, our method could be trained on structural binding sites assigned to proteins in the work of Kim *et al.* [136] and then use it to assign binding sites to proteins for which no structural information is available.

6.1.7 PIANA and Diseases

Protein interaction networks have been shown to be a useful approach to characterizing diseases caused by malfunctions in genes or proteins [130, 145, 147]. In this thesis, we used PIANA to predict cancer genes by integrating protein interaction networks, differential expression studies and structural, functional and evolutionary properties of genes and proteins. However, this integration required manual collection of expression data [216], and the application of PIANA to study other diseases (or cancer experiments) would require the definition of (i) what is considered a disease; (ii) which data is available for the disease of interest; and (iii) interpreting the disease based on the available data. Future work on PIANA should be directed towards automatically obtaining disease profiles and analyzing them in combination with molecular interaction data. For example, the PIANA database could be extended to contain (i) differential expression studies; (ii) gene copy-number alterations; (iii) co-expression data; (iii) third-party repositories of disease information such as the Mendelian Inheritance in Man (MIM) [228]; (iv) databases linking compounds and their target biomolecules [229]; and (v) databases with diseases and drugs used to cure them [230]. Ideally, PIANA would receive as input the disease of interest and all information would be combined to (i) provide an overview of the biological processes behind the disease; (ii) predict new proteins related to the disease; and (iii) propose drugs that activate/inhibit the disease proteins.

6.1.8 Using PIANA to detect remote homologs

The first article included in Chapter V describes and evaluates a method that uses both sequence similarity and protein interactions to identify relationships between remotely related protein sequences [231]. This method relies on the fact that homologous sequences share similar interactions, and therefore, the set of interacting partners of the partners of a

given protein is enriched by its homologs. The results presented in by Espadaler, Aragues *et al.* [231] were based on interactions from the Database of Interacting Proteins (DIP) [105]. However, DIP does not contain all protein interaction data that is publicly available, and therefore, the method would benefit from the use of a larger number of interactions. We are currently developing a web server for the method described by Espadaler, Aragues *et al.* [231] which will access to interactions retrieved by PIANA from DIP [105], MIPS [108], HPRD [232], BIND [107], IntAct. [104], BioGrid [110] and MINT [109]. We expect that the use of PIANA will augment the sensitivity and the applicability of the method and thus, a larger number of remote homologs will be detected for most proteins.

6.1.9 The path is consensus, not integration

A prerequisite to computational biology is the integration of heterogeneous experimental data, which are stored in numerous life-science databases. However, a wide range of obstacles that relate to access, handling and integration impede the efficient use of the contents of these databases [158]. Consequently and coherently, a large part of the work presented in this thesis has consisted in parsing and integrating data coming from multiple and heterogeneous sources. We have achieved a certain success at the integration of protein interaction data, and we are now able to work with an integrated network that contains 405,808 interactions between 53,143 different proteins. However, the path towards optimal integration is not the use of integrative methods but rather building a consensus between the major players in the field. Recent years have seen an enormous activity in data standardization initiatives [74, 75, 169], nomenclature systems [163-166], literature markup languages [233] and controlled vocabularies [64]. However, these initiatives won't be successful without substantial efforts dedicated to educating the research community in the use of standards and thus, their practical implementation and frequency of use is still far from optimal. Hopefully, future computational biology researchers will be able to dedicate most of their time to research, and parsing and integration will become antique concepts. The path to integration is consensus, not integration.

CONCLUSIONS

This section describes in short the main achievements of the work presented in this thesis.

1. A new software framework for integrating, managing and analyzing protein-protein interactions has been created from scratch. This software, PIANA, has been adopted as a work platform by most members of my laboratory, and externally in laboratories around the world.

2. PIANA has been used to unify most public repositories of protein interactions into a single database. The analysis of interactions in this database has been used to assess the protein interaction data available in the public domain. The low overlap found between different sources of protein interaction data demonstrates the need for integrative methods that unify all interactions into a single network.

3. A method for detecting groups of proteins with a common interacting motif has been described. The results from this method suggest that two properties of protein hubs previously attributed to their large number of interactions, (i) the high likelihood of being essential and (ii) the slow evolutionary rate, are actually more reflective of the number of interacting motifs in the protein hub.

4. A method for predicting cancer gene candidates has been described. The integration of protein interaction networks, differential expression data and structural, functional and evolutionary properties of cancer genes has been shown to outperform methods that rely on one single source of data.

5. Protein-protein interactions have been shown to be useful at analyzing cancer and proposing potential cancer genes. In particular, predictions from our integrative approach have been shown to be reliable, and other cancer gene candidates have been validated in collaboration with wet lab experimentalists.

6. In addition to the bioinformatics methods presented here, our work is a first step towards bridging the gap between experimentalist biologists and computational biologists. This has been achieved by creating a tool that has demonstrated its capabilities in contexts ranging from the answer of basic biological questions to the prioritization of cancer gene candidates.

EPILOGUE

Coming from a computational science background, there was only one thing I could think of during my first months of bioinformatics sufferings: what are these people doing?

In my previous life, I was used to more-or-less stable data models, well formatted input data, good understanding of mechanisms behind the subject of study and robust and flexible software. However, I was facing a world in which not much was known about the subject of study (i.e. biology), where data models changed overnight, data seemed to be formatted with the objective of maximizing the errors when parsing and, in most cases, software was a collection of scripts made on the fly.

Somebody said that in order to beat your enemy you have to become like him. Now, I am proud to announce that I have become one of them: I produce my own each-release-is-different database; my software is plagued with unorthodox pieces of code; and I don't aspire to understand things before developing a hypothesis.

Now, on the serious side –was I not serious previously?- there is one thing that not that many people in bioinformatics pay attention to, but that I would guess is as important as having a good understanding of biology: organization. Let me complain a bit: it doesn't make sense that each bioinformatician in the world has his own parser for each database he utilizes; It doesn't make sense that databases are not coherent even with respect to themselves; It doesn't make sense that genes and proteins are identified by non-unique names that follow a "many to many" relationship with reality; It doesn't make sense that a biologist starting a PhD in bioinformatics spends the first year of his PhD learning how to (incorrectly) parse data files; and it doesn't make sense that most bioinformaticians, including myself, apply methods on the basis that "everyone does it". There are many things in bioinformatics that do not make sense. However, bioinformatics has brought great advancements to the understanding of the biology of the cell. Imagine what would had happened (and will happen) if bioinformatics started making some more sense…

Maybe I am writing this because I have become a scientist. Maybe I am just finishing my PhD and there is little energy left in my body. Maybe everyone in the world will read this and bioinformatics will become an organized science. Maybe not.

INDEX

THESIS REFERENCES

1. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
2. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.
3. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
4. Li, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
5. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.
6. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.
7. Cusick, M.E., et al., *Interactome: gateway into systems biology.* Hum Mol Genet, 2005. **14 Spec No. 2**: p. R171-81.
8. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
9. Lodish, H.F., *Molecular cell biology*. 5th ed. ed. 2003, New York: W. H. Freeman ; Basingstoke : [Palgrave]. xxxiii, 973 , [79] p.
10. Aloy, P. and R.B. Russell, *Structural systems biology: modelling protein interactions.* Nature reviews, 2006. **7**(3): p. 188-97.
11. Hoffmann, R. and A. Valencia, *A gene network for navigating the literature.* Nature genetics, 2004. **36**(7): p. 664.
12. *Bioinformatics Definition Committee, NIH working definition of bioinformatics and computational biology.* www.bisti.nih.gov/CompuBioDef.pdf, 2000.
13. Kahn, P., *From genome to proteome: looking at a cell's proteins.* Science, 1995. **270**(5235): p. 369-70.
14. Dongre, A.R., J.K. Eng, and J.R. Yates, 3rd, *Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins.* Trends Biotechnol, 1997. **15**(10): p. 418-25.
15. Pandey, A. and M. Mann, *Proteomics to study genes and genomes.* Nature, 2000. **405**(6788): p. 837-46.
16. Clore, G.M., et al., *A comparison of the restrained molecular dynamics and distance geometry methods for determining three-dimensional structures of proteins on the basis of interproton distances.* FEBS Lett, 1987. **213**(2): p. 269-77.
17. Wuthrich, K., *Protein structure determination in solution by NMR spectroscopy.* J Biol Chem, 1990. **265**(36): p. 22059-62.
18. Wagner, G., *An account of NMR in structural biology.* Nat Struct Biol, 1997. **4 Suppl**: p. 841-4.
19. Brenner, S.E., *A tour of structural genomics.* Nat Rev Genet, 2001. **2**(10): p. 801-9.
20. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.
21. Schulze, A. and J. Downward, *Navigating gene expression using microarrays--a technology review.* Nat Cell Biol, 2001. **3**(8): p. E190-5.
22. Zhu, H., et al., *Global analysis of protein activities using proteome chips.* Science, 2001. **293**(5537): p. 2101-5.

23. Chen, C.S. and H. Zhu, *Protein microarrays.* Biotechniques, 2006. **40**(4): p. 423, 425, 427 passim.
24. Ptacek, J., et al., *Global analysis of protein phosphorylation in yeast.* Nature, 2005. **438**(7068): p. 679-84.
25. Buck, M.J. and J.D. Lieb, *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.* Genomics, 2004. **83**(3): p. 349-60.
26. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotechnol, 2005. **23**(1): p. 137-44.
27. Kumar, A., et al., *Subcellular localization of the yeast proteome.* Genes Dev, 2002. **16**(6): p. 707-19.
28. Huh, W.K., et al., *Global analysis of protein localization in budding yeast.* Nature, 2003. **425**(6959): p. 686-91.
29. Phizicky, E.M. and S. Fields, *Protein-protein interactions: methods for detection and analysis.* Microbiol Rev, 1995. **59**(1): p. 94-123.
30. Estojak, J., R. Brent, and E.A. Golemis, *Correlation of two-hybrid affinity data with in vitro measurements.* Mol Cell Biol, 1995. **15**(10): p. 5820-9.
31. Van Criekinge, W. and R. Beyaert, *Yeast Two-Hybrid: State of the Art.* Biol Proced Online, 1999. **2**: p. 1-38.
32. Parrish, J.R., K.D. Gulyas, and R.L. Finley, Jr., *Yeast two-hybrid contributions to interactome mapping.* Curr Opin Biotechnol, 2006. **17**(4): p. 387-93.
33. Puig, O., et al., *The tandem affinity purification (TAP) method: a general procedure of protein complex purification.* Methods, 2001. **24**(3): p. 218-29.
34. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.
35. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.
36. Ewing, R.M., et al., *Large-scale mapping of human protein-protein interactions by mass spectrometry.* Mol Syst Biol, 2007. **3**: p. 89.
37. Jones, R.B., et al., *A quantitative protein interaction network for the ErbB receptors using protein microarrays.* Nature, 2006. **439**(7073): p. 168-74.
38. Barrios-Rodiles, M., et al., *High-throughput mapping of a dynamic signaling network in mammalian cells.* Science, 2005. **307**(5715): p. 1621-5.
39. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.
40. Serebriiskii, I.G., V. Khazak, and E.A. Golemis, *Redefinition of the yeast two-hybrid system in dialogue with changing priorities in biological research.* BioTechniques, 2001. **30**(3): p. 634-6, 638, 640 passim.
41. Bader, G.D. and C.W. Hogue, *Analyzing yeast protein-protein interaction data obtained from different sources.* Nature biotechnology, 2002. **20**(10): p. 991-7.
42. Kuroda, K., et al., *Systems for the detection and analysis of protein-protein interactions.* Applied microbiology and biotechnology, 2006. **71**(2): p. 127-36.
43. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
44. Sprinzak, E., S. Sattath, and H. Margalit, *How reliable are experimental protein-protein interaction data?* J Mol Biol, 2003. **327**(5): p. 919-23.
45. Stelzl, U. and E.E. Wanker, *The value of high quality protein-protein interaction networks for systems biology.* Curr Opin Chem Biol, 2006. **10**(6): p. 551-8.
46. Valencia, A. and F. Pazos, *Computational methods for the prediction of protein interactions.* Curr Opin Struct Biol, 2002. **12**(3): p. 368-73.

47. Salwinski, L. and D. Eisenberg, *Computational methods of analysis of protein-protein interactions.* Curr Opin Struct Biol, 2003. **13**(3): p. 377-82.
48. Shoemaker, B.A. and A.R. Panchenko, *Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.* PLoS computational biology, 2007. **3**(4): p. e43.
49. Matthews, L.R., et al., *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".* Genome Res, 2001. **11**(12): p. 2120-6.
50. Pazos, F. and A. Valencia, *In silico two-hybrid system for the selection of physically interacting protein pairs.* Proteins, 2002. **47**(2): p. 219-27.
51. Espadaler, J., et al., *Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships.* Bioinformatics, 2005. **21**(16): p. 3360-8.
52. Sprinzak, E. and H. Margalit, *Correlated sequence-signatures as markers of protein-protein interaction.* J Mol Biol, 2001. **311**(4): p. 681-92.
53. Enright, A.J., et al., *Protein interaction maps for complete genomes based on gene fusion events.* Nature, 1999. **402**(6757): p. 86-90.
54. Krallinger, M. and A. Valencia, *Text-mining and information-retrieval services for molecular biology.* Genome Biol, 2005. **6**(7): p. 224.
55. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.
56. Marcotte, E.M., et al., *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. **285**(5428): p. 751-3.
57. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.
58. Suthram, S., et al., *A direct comparison of protein interaction confidence assignment schemes.* BMC bioinformatics, 2006. **7**: p. 360.
59. Deane, C.M., et al., *Protein interactions: two methods for assessment of the reliability of high throughput observations.* Molecular & cellular proteomics, 2002. **1**(5): p. 349-56.
60. Deng, M., F. Sun, and T. Chen, *Assessment of the reliability of protein-protein interactions and protein function prediction.* Pacific Symposium on Biocomputing, 2003: p. 140-51.
61. Saito, R., H. Suzuki, and Y. Hayashizaki, *Interaction generality, a measurement to assess the reliability of a protein-protein interaction.* Nucleic acids research, 2002. **30**(5): p. 1163-8.
62. Chen, J., et al., *Increasing confidence of protein interactomes using network topological metrics.* Bioinformatics (Oxford, England), 2006. **22**(16): p. 1998-2004.
63. Goldberg, D.S. and F.P. Roth, *Assessing experimentally derived interactions in a small world.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(8): p. 4372-6.
64. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
65. Patil, A. and H. Nakamura, *Filtering high-throughput protein-protein interaction data using a combination of genomic features.* BMC bioinformatics, 2005. **6**: p. 100.
66. Bader, J.S., et al., *Gaining confidence in high-throughput protein interaction networks.* Nat Biotechnol, 2004. **22**(1): p. 78-85.
67. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

68. Schierz, A.C., L.N. Soldatova, and R.D. King, *Overhauling the PDB.* Nat Biotechnol, 2007. **25**(4): p. 437-42.
69. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2007. **35**(Database issue): p. D26-31.
70. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D187-91.
71. Pearson, H., *Biology's name game.* Nature, 2001. **411**(6838): p. 631-2.
72. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-5.
73. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): p. D747-50.
74. Ball, C.A., et al., *Submission of microarray data to public repositories.* PLoS Biol, 2004. **2**(9): p. E317.
75. Spellman, P.T., et al., *Design and implementation of microarray gene expression markup language (MAGE-ML).* Genome Biol, 2002. **3**(9): p. RESEARCH0046.
76. Chothia, C., *Proteins. One thousand families for the molecular biologist.* Nature, 1992. **357**(6379): p. 543-4.
77. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
78. Liu, J. and B. Rost, *Domains, motifs and clusters in the protein universe.* Curr Opin Chem Biol, 2003. **7**(1): p. 5-11.
79. Pearl, F., et al., *The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis.* Nucleic Acids Res, 2005. **33**(Database issue): p. D247-51.
80. Brenner, S.E., P. Koehl, and M. Levitt, *The ASTRAL compendium for protein structure and sequence analysis.* Nucleic Acids Res, 2000. **28**(1): p. 254-6.
81. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.
82. Mulder, N.J., et al., *InterPro, progress and status in 2005.* Nucleic Acids Res, 2005. **33**(Database issue): p. D201-5.
83. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D227-30.
84. Attwood, T.K., et al., *PRINTS and its automatic supplement, prePRINTS.* Nucleic Acids Res, 2003. **31**(1): p. 400-2.
85. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* J Mol Biol, 2001. **313**(4): p. 903-19.
86. Servant, F., et al., *ProDom: automated clustering of homologous domains.* Brief Bioinform, 2002. **3**(3): p. 246-51.
87. Murali, T.M., C.J. Wu, and S. Kasif, *The art of gene function prediction.* Nature biotechnology, 2006. **24**(12): p. 1474-5; author reply 1475-6.
88. Sharan, R., I. Ulitsky, and R. Shamir, *Network-based prediction of protein function.* Mol Syst Biol, 2007. **3**: p. 88.
89. Ouzounis, C.A., et al., *Classification schemes for protein structure and function.* Nature reviews, 2003. **4**(7): p. 508-19.
90. Bairoch, A., *The ENZYME database in 2000.* Nucleic acids research, 2000. **28**(1): p. 304-5.
91. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

92. Bader, G.D., M.P. Cary, and C. Sander, *Pathguide: a pathway resource list.* Nucleic acids research, 2006. **34**(Database issue): p. D504-6.
93. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic acids research, 2006. **34**(Database issue): p. D354-7.
94. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic Acids Res, 2005. **33**(Database issue): p. D428-32.
95. Dahlquist, K.D., et al., *GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.* Nature genetics, 2002. **31**(1): p. 19-20.
96. Kunin, V., et al., *Myriads of protein families, and still counting.* Genome biology, 2003. **4**(2): p. 401.
97. Carpenter, A.E. and D.M. Sabatini, *Systematic genome-wide screens of gene function.* Nature reviews, 2004. **5**(1): p. 11-22.
98. Whisstock, J.C. and A.M. Lesk, *Prediction of protein function from protein sequence and structure.* Quarterly reviews of biophysics, 2003. **36**(3): p. 307-40.
99. Heger, A. and L. Holm, *Towards a covering set of protein family profiles.* Prog Biophys Mol Biol, 2000. **73**(5): p. 321-37.
100. Bartlett, G.J., A.E. Todd, and J.M. Thornton, *Inferring protein function from structure.* Methods of biochemical analysis, 2003. **44**: p. 387-407.
101. Lee, I., et al., *A probabilistic functional network of yeast genes.* Science (New York, N.Y, 2004. **306**(5701): p. 1555-8.
102. Mathivanan, S., et al., *An evaluation of human protein-protein interaction data in the public domain.* BMC Bioinformatics, 2006. **7 Suppl 5**: p. S19.
103. Shoemaker, B.A. and A.R. Panchenko, *Deciphering Protein-Protein Interactions. Part I. Experimental Techniques and Databases.* 2007. **3**(3): p. e42.
104. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.
105. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
106. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
107. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33**(Database issue): p. D418-24.
108. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-4.
109. Chatr-aryamontri, A., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
110. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
111. Guldener, U., et al., *CYGD: the Comprehensive Yeast Genome Database.* Nucleic acids research, 2005. **33**(Database issue): p. D364-8.
112. von Mering, C., et al., *STRING 7--recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.
113. Persico, M., et al., *HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.* BMC Bioinformatics, 2005. **6 Suppl 4**: p. S21.
114. Han, K., et al., *HPID: the Human Protein Interaction Database.* Bioinformatics, 2004. **20**(15): p. 2466-70.

115. Davis, F.P. and A. Sali, *PIBASE: a comprehensive database of structurally defined protein interfaces.* Bioinformatics (Oxford, England), 2005. **21**(9): p. 1901-7.
116. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure (London, England, 1997. **5**(8): p. 1093-108.
117. Winter, C., et al., *SCOPPI: a structural classification of protein-protein interfaces.* Nucleic acids research, 2006. **34**(Database issue): p. D310-4.
118. Finn, R.D., M. Marshall, and A. Bateman, *iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.* Bioinformatics (Oxford, England), 2005. **21**(3): p. 410-2.
119. Deng, M., et al., *Inferring domain-domain interactions from protein-protein interactions.* Genome Res, 2002. **12**(10): p. 1540-8.
120. Ng, S.K., Z. Zhang, and S.H. Tan, *Integrative approach for computationally inferring protein domain interactions.* Bioinformatics, 2003. **19**(8): p. 923-9.
121. Ohkubo, J., M. Yasuda, and K. Tanaka, *Statistical-mechanical iterative algorithms on complex networks.* Phys Rev E Stat Nonlin Soft Matter Phys, 2005. **72**(4 Pt 2): p. 046135.
122. Almaas, E., *Biological impacts and context of network theory.* J Exp Biol, 2007. **210**(Pt 9): p. 1548-58.
123. Kwoh, C.K. and P.Y. Ng, *Network analysis approach for biology.* Cell Mol Life Sci, 2007.
124. Davidson, E.H., et al., *A genomic regulatory network for development.* Science, 2002. **295**(5560): p. 1669-78.
125. Sambrano, G.R., et al., *Unravelling the signal-transduction network in B lymphocytes.* Nature, 2002. **420**(6916): p. 708-10.
126. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-4.
127. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.
128. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
129. Butte, A.J. and I.S. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. **24**(1): p. 55-62.
130. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.
131. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science (New York, N.Y, 2003. **302**(5651): p. 1727-36.
132. Yook, S.H., Z.N. Oltvai, and A.L. Barabasi, *Functional and topological characterization of protein interaction networks.* Proteomics, 2004. **4**(4): p. 928-42.
133. Rachlin, J., et al., *Biological context networks: a mosaic view of the interactome.* Mol Syst Biol, 2006. **2**: p. 66.
134. Han, J.D., et al., *Effect of sampling on topology predictions of protein-protein interaction networks.* Nat Biotechnol, 2005. **23**(7): p. 839-44.
135. Tanaka, R., T.M. Yi, and J. Doyle, *Some protein interaction data do not exhibit power law statistics.* FEBS Lett, 2005. **579**(23): p. 5140-4.
136. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.
137. Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.
138. Yu, H., et al., *The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics.* PLoS Comput Biol, 2007. **3**(4): p. e59.

139. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.
140. Gagneur, J., et al., *Modular decomposition of protein-protein interaction networks.* Genome Biol, 2004. **5**(8): p. R57.
141. Spirin, V. and L.A. Mirny, *Protein complexes and functional modules in molecular networks.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12123-8.
142. Hu, Z., et al., *Towards zoomable multidimensional maps of the cell.* Nature biotechnology, 2007. **25**(5): p. 547-54.
143. Guimaraes, K.S., et al., *Predicting domain-domain interactions using a parsimony approach.* Genome Biol, 2006. **7**(11): p. R104.
144. Lehner, B. and A.G. Fraser, *A first-draft human protein-interaction map.* Genome Biol, 2004. **5**(9): p. R63.
145. Rhodes, D.R., et al., *Probabilistic model of the human protein-protein interaction network.* Nat Biotechnol, 2005. **23**(8): p. 951-9.
146. Milo, R., et al., *Superfamilies of evolved and designed networks.* Science, 2004. **303**(5663): p. 1538-42.
147. Gandhi, T.K., et al., *Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.* Nat Genet, 2006. **38**(3): p. 285-93.
148. Huynen, M.A., et al., *Function prediction and protein networks.* Curr Opin Cell Biol, 2003. **15**(2): p. 191-8.
149. Chua, H.N., W.K. Sung, and L. Wong, *Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions.* Bioinformatics, 2006. **22**(13): p. 1623-30.
150. Nabieva, E., et al., *Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps.* Bioinformatics, 2005. **21 Suppl 1**: p. i302-10.
151. Deng, M., et al., *Mapping Gene Ontology to proteins based on protein-protein interaction data.* Bioinformatics, 2004. **20**(6): p. 895-902.
152. Samanta, M.P. and S. Liang, *Predicting protein functions from redundancies in large-scale protein interaction networks.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12579-83.
153. Albert, I. and R. Albert, *Conserved network motifs allow protein-protein interaction prediction.* Bioinformatics (Oxford, England), 2004. **20**(18): p. 3346-52.
154. Stein, L.D., *Integrating biological databases.* Nature reviews, 2003. **4**(5): p. 337-45.
155. Hwang, D., et al., *A data integration methodology for systems biology.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(48): p. 17296-301.
156. Shah, A.R., et al., *Enabling high-throughput data management for systems biology: The Bioinformatics Resource Manager.* 2007.
157. Searls, D.B., *Data integration: challenges for drug discovery.* Nature reviews, 2005. **4**(1): p. 45-58.
158. Philippi, S. and J. Kohler, *Addressing the problems with life-science databases for traditional uses and systems biology.* Nature reviews, 2006. **7**(6): p. 482-8.
159. Davidson, S.B., et al., *K2/Kleisli and GUS: Experiments in integrated access to genomic data sources.* IBM SYSTEMS JOURNAL, 2001. **40**(2).
160. Koehler, J., et al., *Linking experimental results, biological networks and sequence analysis methods using Ontologies and Generalised Data Structures.* In silico biology, 2005. **5**(1): p. 33-44.
161. Draghici, S., S. Sellamuthu, and P. Khatri, *Babel's tower revisited: a universal resource for cross-referencing across annotation databases.* Bioinformatics (Oxford, England), 2006. **22**(23): p. 2934-9.

162. Olson, A.J., T. Tully, and R. Sachidanandam, *GeneSeer: a sage for gene names and genomic resources.* BMC genomics, 2005. **6**: p. 134.
163. Wain, H.M., et al., *Guidelines for human gene nomenclature.* Genomics, 2002. **79**(4): p. 464-70.
164. Crosby, M.A., et al., *FlyBase: genomes by the dozen.* Nucleic acids research, 2007. **35**(Database issue): p. D486-91.
165. Bieri, T., et al., *WormBase: new content and better access.* Nucleic acids research, 2007. **35**(Database issue): p. D506-10.
166. Cherry, J.M., *Genetic nomenclature guide. Saccharomyces cerevisiae.* Trends in genetics, 1995: p. 11-2.
167. *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2007. **35**(Database issue): p. D193-7.
168. Hubbard, T.J., et al., *Ensembl 2007.* Nucleic acids research, 2007. **35**(Database issue): p. D610-7.
169. Hermjakob, H., *The HUPO Proteomics Standards Initiative - Overcoming the Fragmentation of Proteomics Data.* Proteomics, 2006. **6 Suppl 2**: p. 34-8.
170. Consortium, I.M.E., http://imex.sourceforge.net/.
171. Hart, G.T., A.K. Ramani, and E.M. Marcotte, *How complete are current yeast and human protein-interaction networks?* Genome Biol, 2006. **7**(11): p. 120.
172. Zhao, R. and W.A. Houry, *Molecular interaction network of the Hsp90 chaperone system.* Advances in experimental medicine and biology, 2007. **594**: p. 27-36.
173. Hwang, D., et al., *A data integration methodology for systems biology: experimental verification.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(48): p. 17302-7.
174. Jansen, R., D. Greenbaum, and M. Gerstein, *Relating whole-genome expression data with protein-protein interactions.* Genome research, 2002. **12**(1): p. 37-46.
175. de Lichtenberg, U., et al., *Dynamic complex formation during the yeast cell cycle.* Science (New York, N.Y, 2005. **307**(5710): p. 724-7.
176. Sandmann, T., et al., *A core transcriptional network for early mesoderm development in Drosophila melanogaster.* Genes & development, 2007. **21**(4): p. 436-49.
177. Tanay, A., et al., *Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(9): p. 2981-6.
178. Bluthgen, N., et al., *HOMGL-comparing genelists across species and with different accession numbers.* Bioinformatics (Oxford, England), 2004. **20**(1): p. 125-6.
179. Khatri, P., et al., *New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate.* Nucleic acids research, 2006. **34**(Web Server issue): p. W626-31.
180. Pillet, V., et al., *GPSDB: a new database for synonyms expansion of gene and protein names.* Bioinformatics (Oxford, England), 2005. **21**(8): p. 1743-4.
181. Bussey, K.J., et al., *MatchMiner: a tool for batch navigation among gene and gene product identifiers.* Genome biology, 2003. **4**(4): p. R27.
182. Liu, H., et al., *BioThesaurus: a web-based thesaurus of protein and gene names.* Bioinformatics (Oxford, England), 2006. **22**(1): p. 103-5.
183. Burdick, D., *Connect the dots.* 2007.
184. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
185. Hu, Z., et al., *VisANT: an online visualization and analysis tool for biological interaction data.* BMC Bioinformatics, 2004. **5**: p. 17.

186. Breitkreutz, B.J., C. Stark, and M. Tyers, *Osprey: a network visualization system.* Genome Biol, 2003. **4**(3): p. R22.
187. Cerami, E.G., et al., *cPath: open source software for collecting, storing, and querying biological pathways.* BMC Bioinformatics, 2006. **7**: p. 497.
188. Iragne, F., et al., *ProViz: protein interaction visualization and exploration.* Bioinformatics, 2005. **21**(2): p. 272-4.
189. Meil, A., P. Durand, and J. Wojcik, *PIMWalker: visualising protein interaction networks using the HUPO PSI molecular interaction format.* Appl Bioinformatics, 2005. **4**(2): p. 137-9.
190. Yip, K.Y., et al., *The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks.* Bioinformatics, 2006. **22**(23): p. 2968-70.
191. Tarassov, K. and S.W. Michnick, *iVici: Interrelational Visualization and Correlation Interface.* Genome Biol, 2005. **6**(13): p. R115.
192. Han, K. and B.H. Ju, *A fast layout algorithm for protein interaction networks.* Bioinformatics, 2003. **19**(15): p. 1882-8.
193. Chang, A.N., et al., *INTEGRATOR: interactive graphical search of large protein interactomes over the Web.* BMC Bioinformatics, 2006. **7**: p. 146.
194. Jiang, T. and A.E. Keating, *AVID: an integrative framework for discovering functional relationships among proteins.* BMC bioinformatics, 2005. **6**: p. 136.
195. Chaurasia, G., et al., *UniHI: an entry gate to the human protein interactome.* Nucleic acids research, 2007. **35**(Database issue): p. D590-4.
196. Martin-Sanchez, F., et al., *Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care.* Journal of biomedical informatics, 2004. **37**(1): p. 30-42.
197. Loging, W., L. Harland, and B. Williams-Jones, *High-throughput electronic biology: mining information for drug discovery.* Nature reviews, 2007. **6**(3): p. 220-30.
198. Thiel, K.A., *Structure-aided drug design's next generation.* Nature biotechnology, 2004. **22**(5): p. 513-9.
199. Beerenwinkel, N., et al., *Computational methods for the design of effective therapies against drug resistant HIV strains.* Bioinformatics (Oxford, England), 2005. **21**(21): p. 3943-50.
200. Segal, E., et al., *From signatures to models: understanding cancer using microarrays.* Nat Genet, 2005. **37 Suppl**: p. S38-45.
201. Quackenbush, J., *Computational approaches to analysis of DNA microarray data.* Methods of information in medicine, 2006. **45 Suppl 1**: p. 91-103.
202. Laaksonen, R., et al., *A systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle.* PLoS ONE, 2006. **1**: p. e97.
203. Hood, L. and R.M. Perlmutter, *The impact of systems approaches on biological problems in drug discovery.* Nature biotechnology, 2004. **22**(10): p. 1215-7.
204. Shoichet, B.K., *Virtual screening of chemical libraries.* Nature, 2004. **432**(7019): p. 862-5.
205. Nicholson, J.K. and I.D. Wilson, *Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism.* Nature reviews, 2003. **2**(8): p. 668-76.
206. Mathew, J.P., et al., *From bytes to bedside: data integration and computational biology for translational cancer research.* PLoS computational biology, 2007. **3**(2): p. e12.

207. Hanauer, D.A., et al., *Bioinformatics approaches in the study of cancer.* Current molecular medicine, 2007. **7**(1): p. 133-41.
208. Cho, W.C., *Contribution of oncoproteomics to cancer biomarker discovery.* Molecular cancer, 2007. **6**: p. 25.
209. Vogelstein, B. and K.W. Kinzler, *Cancer genes and the pathways they control.* Nat Med, 2004. **10**(8): p. 789-99.
210. Hu, P., et al., *Computational prediction of cancer-gene function.* Nat Rev Cancer, 2007. **7**(1): p. 23-34.
211. Chung, C.H., et al., *Genomics and proteomics: emerging technologies in clinical cancer research.* Critical reviews in oncology/hematology, 2007. **61**(1): p. 1-25.
212. Rhodes, D.R., et al., *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9309-14.
213. Segal, E., et al., *A module map showing conditional activity of expression modules in cancer.* Nat Genet, 2004. **36**(10): p. 1090-8.
214. Jonsson, P.F. and P.A. Bates, *Global topological features of cancer proteins in the human interactome.* Bioinformatics (Oxford, England), 2006. **22**(18): p. 2291-7.
215. Rhodes, D.R. and A.M. Chinnaiyan, *Integrative analysis of the cancer transcriptome.* Nat Genet, 2005. **37 Suppl**: p. S31-7.
216. Rhodes, D.R., et al., *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.
217. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.
218. Aittokallio, T. and B. Schwikowski, *Graph-based methods for analysing networks in cell biology.* Briefings in bioinformatics, 2006. **7**(3): p. 243-55.
219. Besemann, C., et al., *BISON: bio-interface for the semi-global analysis of network patterns.* 2006. **1**(1): p. 8.
220. Al-Shahrour, F., et al., *From genes to functional classes in the study of biological systems.* BMC bioinformatics, 2007. **8**: p. 114.
221. Al-Shahrour, F., et al., *FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.* 2007.
222. Patricia García-Gomís, R.A., *Diseño de interface para PIANA.* Projecte fi de carrera (Universitat Pompeu Fabra. Enginyeria Tècnica en Informàtica de Sistemes), 2005. **QA75.5 .P76 2005 v. 36**.
223. Pablo Boixeda, R.A., *Clustering jerárquico de una red de interacción entre proteinas.* Projecte fi de carrera (Universitat Pompeu Fabra. Enginyeria Tècnica en Informàtica de Sistemes), 2005. **QA75.5 .P76 2005 v. 13**.
224. Espana, L., et al., *Bcl-x(L)-mediated changes in metabolic pathways of breast cancer cells: from survival in the blood stream to organ-specific metastasis.* The American journal of pathology, 2005. **167**(4): p. 1125-37.
225. Mendez, O., et al., *Underexpression of transcriptional regulators is common in metastatic breast cancer cells overexpressing Bcl-xL.* Carcinogenesis, 2006. **27**(6): p. 1169-79.
226. Kim, W.K., et al., *The many faces of protein-protein interactions: A compendium of interface geometry.* PLoS Comput Biol, 2006. **2**(9): p. e124.
227. Teyra, J., et al., *SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces.* BMC bioinformatics, 2006. **7**: p. 104.

228. McKusick, V.A., *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders.* . Baltimore: Johns Hopkins University Press, 1998 (12th edition). 1998.
229. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.* Nucleic acids research, 2007. **35**(Database issue): p. D198-201.
230. Chen, X., Z.L. Ji, and Y.Z. Chen, *TTD: Therapeutic Target Database.* Nucleic acids research, 2002. **30**(1): p. 412-5.
231. Espadaler, J., et al., *Detecting remotely related proteins by their interactions and sequence similarity.* Proc Natl Acad Sci U S A, 2005. **102**(20): p. 7151-6.
232. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic Acids Res, 2004. **32**(Database issue): p. D497-501.
233. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic acids research, 2004. **32**(Database issue): p. D267-70.

NOTES