Universitá di Roma Tor Vergata

Facoltá di Scienze Matematiche Fisiche e Naturali

Dipartimento di Biologia

Dottorato di Ricerca in Sistemi Complessi Applicati alla Biologia Post-Genomica

CICLO XX

COMPUTATIONAL FRAMEWORKS FOR WIRING HUMAN AND YEAST PROTEOMES

TESI PRESENTATA DA:
Dott.ssa Maria Persico

TUTOR:
Prof. R. Calogero

COORDINATORE DEL CICLO: Prof. F. Bussolino
RELATORE ESTERNO: Prof. G. Cesareni

Anni Accademici: 2004 - 2007
SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA: BIOL-

To my family

Contents

Abstract **7**

Acknowledgements **9**

1 Biological background **11**

1.1 Experimental methods for identifying and characterizing protein interactions 11

1.2 Protein and domain interaction databases. 15

2 Bioinformatics background **17**

2.1 Protein networks as graphs 17

2.1.1 Basic topological properties 17

2.2 Machine learning frameworks 18

2.2.1 Supervised classification 19

2.3 Information Theory to extract knowledge from multiple sequence alignments 21

3 Proposed frameworks for the proteome wiring problem **25**

3.1 Human proteome wiring by using interologs in model organisms: the HomoMINT interactome 25

3.1.1 Data Sources 26

3.1.2 Software 27

3.1.3 Assembly of the orthology table 27

3.1.4 Assembling HEN (Human Experimental Network) 27

3.1.5 Filtering orthology groups for domain architecture homogeneity 28

3.1.6 Gene Ontology similarity analysis 29

3.2 Wiring of proteome in yeast by data integration 30

3.2.1 Data sources 31

3.2.2 Software 31

3.2.3 Assembling the gold standard 32

3.3 Identification of protein interactions in human by co-evolutionary information of the interacting partners . . . 33
3.3.1 Data Sources . . . 34
3.3.2 Software . . . 35
3.3.3 Assembly of the *SH2ome*, a set of human interacting proteins in which at least one in every interacting protein pair is an *SH2 domain containig* protein . . . 35
3.3.4 Criteria for the assembling of the orthology sets . . . 35

4 Results and discussion **37**
4.1 Results and discussion for the reconstruction of human protein interolog network using evolutionary conserved networks: the HomoMINT interactome . . . 37
4.1.1 Intersection of HomoMINT with the Human experimental network . . . 38
4.1.2 Intersection of HomoMINT with the iHOP resource . . . 39
4.1.3 Interacting proteins sharing GO terms . . . 41
4.1.4 HomoMINT as a graph . . . 43
4.1.5 HomoMINT as a web server . . . 46
4.1.6 Discussion . . . 47
4.2 Results and discussion for the wiring of the yeast proteome by data integration approaches . . . 49
4.2.1 A statistical framework for experimental technique independent high-throughput data assembling after a *data integration based* estimation of protein interaction . . . 49
4.2.2 Building a *data integration based* graph for protein interaction by modeling functional genomic and annotations evidences as predictors of protein interaction. . . . 50
4.2.3 Discussion . . . 52
4.3 Results and discussion for the wiring of the human proteome by evolutionary constraints . . . 54
4.3.1 Correlated evolution analysis . . . 54
4.3.2 Definition of the subinteractome to evaluate in term of evolutionary constraints . . . 55
4.3.3 Correlated evolution analysis of the *Human SH2ome* . . . 55
4.3.4 Definition and implementation of a null model for correlated evolution analysis . . . 56
4.3.5 Protein evolutionary covariation analysis . . . 61

5 Overall discussion and perspectives **65**

Conclusion **67**

A Biological glossary **69**

B Bioinformatic glossary **71**

C Publications **75**

Bibliography **76**

Abstract

One of the major challenges of modern system biology is to decipher how the information for the life processes is encoded in the protein networks of a complex organism. Interactions among proteins serve as an important basis for the biological complexity of higher organisms. In recent years, there have been several large scale efforts to map protein interactions on model organisms.

In this thesis, I address the problem of how to wire the proteomes, an important question, central in systems biology; in human (*H. sapiens*), the difficulties in generating protein interaction data have stimulated the development of sequence based prediction frameworks, i.e. co-evolutionary information of the interacting partners and interologs; following this trend, some pipelines were developed aimed to transfer interaction data from model organism to human (the HomoMINT interactome) and to looking for correlations in the distance matrices representing the trees of the ortholog groups to which the human reference proteins under analysis belong. A third method based on the identification of co-evolving residues displaying statistically significant patterns of co-evolution, as measured by mutual information metric was tested; in yeast (*S. cerevisiae*) , the questions related to the wiring problem remain still unanswred in spite of the abundance of protein interaction data from high-throughput experiments. Unfortunately, these large-scale studies show embarrassing discrepancies in their results and coverage. The recent completion of a comprehensive literature curation effort, have made available an interesting new reference set and stimulated building of a simple logistic regression model on wiring of the yeast proteome, based on the definition of some predictors of functional relationships for the pairs of interacting proteins in the reference set: the probability of sharing a path on the Gene Ontology trees, the degree of correlated evolution, the degree of co-expression and co-abundance. Moreover, the value distributions for the analysed genomic features differ respect the distributions of the same predictors computed in previously defined null models (i.e. artificial protein networks).

The model was evaluated by standard criteria and ROC curve analysis. The complete frameworks were implemented in a suite of R - PERL programs.

Acknowledgements

Below I wish to thank some of the many people that have contributed to this work. It has been a pleasure to work and do research with you.

- I am deeply grateful to open source and free software community. Almost no part of this work could have been done without the informatics and bioinformatics resources provided by these people. Thank you also for givin me their way of thinking.
- Thanks to Susana Bueno and Giovanni Chillemi in CASPUR (CONSORZIO INTERUNIVERSITARIO PER LE APPLICAZIONI DI SUPERCALCOLO PER UNIVERSIT E RICERCA) for their friendly informatics support during these years.
- Thanks to the statistician and something more Gianpaolo Scalia Tomba (Rome, University of Tor Vergata)
- Thanks to my husband Rocco

Chapter 1

Biological background

Proteins rarely act alone; rather, they carry out their activities through a multitude of interactions with other proteins or molecules, either pairwise or in association with multiple subunits, as is the case for protein complexes. To unravel the global picture of protein interactions in the cell, different experimetal techniques have been developed.

1.1 Experimental methods for identifying and characterizing protein interactions

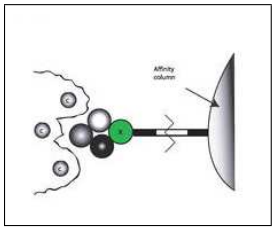
Different genetic, biochemical, and physical methods can be employed to analyze protein interactions: the most used are listed in the table below:

Method	HT	L. c. a.	Int. type	Char. type
Y2H [55, 56]	+	In vivo	Phys. int.(bin)	id.
Affinity purification-MS [10]	+	In vitro	Phys. int.(compl.)	id.
Protein microarrays [**?**, 17, 19]	+	In vitro	Phys. int.(compl.)	id.
synthetic lethality [20, 21]	+	In vivo	Func. ass.	id.
phage display [22]	+	In vitro	Phys. int.(compl.)	id.
X-ray	-	In vitro	Phys. int.(compl.)	Struct. biol char.
fluor. res. energy transfer (FRET) [23]	-	In vivo	Phys. int.(bin)	
surface plasmon resonance (SPR) [24]	-	In vitro	Phys. int.(compl.)	Kin.dyn. char.
Atomic force microscopy [25]	-	In vitro	Phys. int.(bin)	Mech.dyn. char.
Electron microscopy [26]	-	In vitro	Phys. int.(compl.)	Struct. biol char.

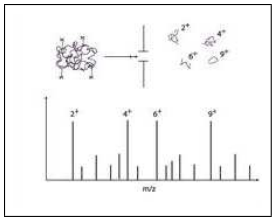
High-throughput techniques are indicated with + symbol (second column), and those which can provide information on interactions in vivo are shown in the third column. Fourth column indicates whether the method supplies data on physically interacting proteins in a complex (complex) or only pairwise

interactions (binary). Methods inferring interactions through functional association are shown as well. The type of protein interaction characterization is shown in the last column.

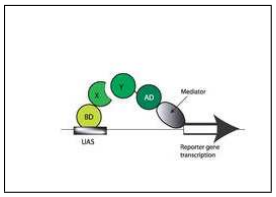
Some techniques enable screening of a large number of proteins in a cell, such as yeast two-hybrid (Y2H), tandem affinity purification (TAP), mass spectroscopy (MS), protein microarrays, synthetic lethality, and phage display. Other methods focus on monitoring and characterizing specific biochemical and physicochemical properties of a protein complex.



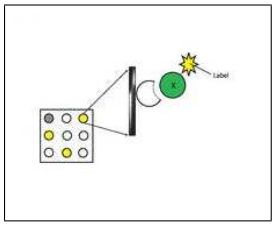
TAP purifies protein complexes and removes the molecules of contaminants.



MS identifies polypeptide sequence.



Y2H detects interactions between proteins X and Y, where X is linked to BD domain which binds to upstream activating sequence (UAS) of a promoter.



In protein microarrays (protein chips) target proteins immobilized on the solid support are probed with a fluorescently labeled protein.

Yeast two-hybrid method.

The development of the Y2H technique has considerably accelerated the screening of protein interactions in vivo. Y2H is based on the fact that many eukaryotic transcription activators have at least two distinct domains, one that directs binding to a promoter DNA sequence (BD) and another that activates transcription (AD) (Figure 1.1). It was demonstrated that splitting BD and AD inactivates the transcription, but the transcription can be restored if a DNA-binding domain is physically (not necessarily covalently) associated with an activating domain [1]. According to the Y2H method, a protein of interest is fused to BD (bait). This chimeric protein is cloned in an expression plasmid, which is then transfected into a yeast cell. A similar procedure creates a chimeric sequence of another protein fused to AD (prey). If two proteins physically interact, the reporter gene is activated. The most broadly used Y2H systems are GAL4/LexA-based, where the GAL4 protein controls in yeast the expression of the LacZ gene encoding beta-galactosidase. Numerous variations of Y2H have been developed including systems with several reporter genes, one-hybrid and three-hybrid systems for identifying proteins interactions with DNA and RNA [2], systems for detecting interactions in mammalian and prokaryotic cells, and systems for screening the interactions between membrane proteins [3, 4].

For screening entire genomes, the Y2H method has been advanced into two main approaches [5, 81] : matrix-based and library-based.

In the matrix approach, a matrix of prey clones is created where each clone expresses a particular prey protein in one well of a plate. Then each bait strain is mated with an array of prey strains and those diploids where two chimeric proteins interact are selected based on the expression of a reporter gene and the position on a plate.

In the library approach, each bait is screened against an undefined prey library containing random cDNA fragments or open reading frames (ORFs). Diploid positives are selected based on their ability to grow on specific substrates; and interacting proteins are determined by DNA sequencing. The first two genome-wide analyses of the yeast interactome revealed 692 and 841 putative interactions, respectively [55, 56]. The overlap between these two experimental studies was quite small; both methods shared only 141 interactions, about 20

The small overlap between Y2H experiments can be explained by different factors, among them: differences in protein interaction sampling, Y2H bias towards nonspecific interactions [6], and limitations of the Y2H method itself. For example, proteins initiating transcription by themselves cannot be targeted in Y2H experiments; and the use of sequence chimeras can impose difficulties since fusion can change the structure of a target protein. In addition, protein folding and posttranslational modifications can differ between yeast and other

organisms. This makes it difficult to screen proteins from mammalian and prokaryotic cells using Y2H as well as cytoplasmic and membrane proteins. To validate the quality of Y2H protein interactions in vivo, different in vitro techniques can be used.

TAP-MS

Several large-scale studies of protein complexes have been performed using TAP MS methods [7, 8]. For example, Krogan et al. showed that 7,123 protein interactions identified with high confidence in yeast can be clustered into 547 protein complexes [9]. The methodology consists in two separated steps:

- TAP method
  A TAP tag consists of two IgG binding domains of Staphylococcus protein A and a calmodulin binding peptide separated by the tobacco etch virus protease cleavage site [10] (Figure 1.1). A target protein open reading frame (ORF) is fused with the DNA sequences encoding the TAP tag and is expressed in yeast where it can form native complexes with other proteins. At the first step of the TAP purification, protein A binds tightly to an IgG matrix; and after washing out the contaminants, the protease cleaves the link between protein A and IgG matrix. The eluate of this first step is then incubated with calmodulin-coated beads in the presence of calcium. After washing, the target protein complex is released. The components of each complex are screened by polyacrylamide gel electrophoresis, cleaved by proteases, and the fragments are identified by MS.

- MS analysis
  MS is a powerful method of studying macromolecular interactions in vitro. The principle of the MS method is to produce ions which can be detected based on their mass-to-charge ratios, thereby allowing the identification of polypeptide sequences [2, 11] (Figure 1.1). The problem of converting protein/peptide molecules from the condensed phase into ions in the gas phase is solved by using Electrospray Ionization (ESI) [12] and Matrix Assisted Laser Desorption Ionization (MALDI) [13]. Different algorithms have been developed to analyze mass spectra and to identify proteins by their sequence [14, 15]. Some of them find correlations between theoretical and experimental spectra while others use de novo algorithms to infer peptide sequences from theoretical interpretation of the mass spectra. Despite the usefulness of MS for the characterization of interacting proteins, purification of protein complexes turns out to be the limiting step of their identification.

Comparing Y2H and TAPMS, it should be noted that both methods generate a lot of false positives and miss a lot of known interactions. Y2H has the advantages of being an in vivo technique and of detecting transient interactions. In contrast, TAPMS can report on higher-order interactions, therefore, provides direct information on protein complexes.

The real-time characterization of interacting proteins in vivo can be achieved with various spectroscopic techniques requiring the attachment of a spectroscopic label to a target protein [16] (Table 1.1). A powerful technique in this respect is fluorescence resonance energy transfer (FRET), which can occur only if two fluorophores are located close to each other [23]. Another effective method, surface plasmon resonance (SPR), does not require spectroscopic labeling and can detect interactions between soluble ligands and immobilized receptors [24]. Recently, new methods have been developed to analyze protein interactions at the single-molecule level. For example, atomic force microscopy can fairly accurately measure interaction forces [25] while fluorescence techniques can characterize conformational changes in proteins upon binding [18].

1.2 Protein and domain interaction databases.

Database	Pr. or Dom.	Type	N. of Int.	URL FTP
DIP a LiveDIP	P	E S	55.733	[27]
BIND a	P	E C S	83.517	[35]
MPact MIPS a	P	E C F	15.488 (4.300)b	[29]
STRING	P	E P F	730.000 (proteins)	[36]
MINT a	P	E C	71.854	[33]
IntAct a	P	E C	68.165	[28]
BioGRID a	P	E C	116.000 (30.000)b	[30]
HPRD	P	E C	33.710	[31]
iPfam	D	S	3.019	[34]
DIMA	D	F S		[32]

Table 1.1: Listed are: the name of the database; the unit of interaction, protein (P) or domain (D); type of data (high-throughput experimental data (E), structural data (S), manual curation (C), functional predictions (F), and interface homology modeling (H)); and the number of interactions. a Databases are members of the International Molecular Exchange Consortium (IMEx) (http://imex.sourceforge.net). b Number of interactions listed in parentheses is for curated set.

A large variety of databases exists storying binary protein interactions and the higher order interactions in protein complexes. A summary of some available databases is given in the Tables 1.2. These resources contain interactions obtained by direct submission from experimentalists or from expert curators mining the literature and other data sources; There is a wide range of detail characterizing the interactions available from different databases even if a standard data model has been proposed for the representation and exchange of protein

interaction data [37]. Due to the interaction data diversity, ranging from large-scale datasets to a single interaction confirmed by several different techniques, often reported either as free text or in tables of variable format, there exists the need to estabilish reporting guidelines avoiding missing key pieces of information essential for a full understanding of the experiment. This concept has been implemented in MIMIx [38], the minimum information required for reporting a molecular interaction experiment. Adherence to these reporting guidelines will result in publications of increased clarity and usefulness to the scientific community and will support the rapid, systematic capture of molecular interaction data in public databases, thereby improving access to valuable interaction data. Another need in this field, is promoting the possibility for each resource to develop own specific tools of analysis of protein interaction data, avoiding the users employing another data base because richer of data; moreover, it is important to deal with potential problem of redundancy; for these reasons, the International Molecular Exchange Consortium (IMEx) [39] has been formed in which databases agree to share their data in a consistent and timely fashion. In conclusion, interacting proteins can be studied either as complete units or by domains used as the units of interaction. Consequently, some databases domain-related have been developed [40–42].

Chapter 2

Bioinformatics background

The experimental methods aimed at charactherizing the proteomes and detecting protein-protein interactions are time-consuming and costly, which has motivated vigorous development of computational approaches to predict the functional links and to wire the proteins in the living systems.

We can define the wiring of the proteomes as the attempt to define and to select the most likely graph among all possible that one can build with a specified set of nodes (proteins).

2.1 Protein networks as graphs

A graph is a collection of nodes and edges connecting those nodes. A *protein protein interaction graph* (or network) G is formally defined by the node set N representing all proteins found within the protein interaction data set and the edge set E that represents the set of physical binding or a functional relationships between the proteins.

2.1.1 Basic topological properties

In this graph, the degree k_i counts the number of interactions that node i has with the other nodes in the network. By averaging k_i over all nodes in the network we obtain the *average degree*.

$$K = \frac{1}{N}\sum_{i=1)}^{N} K(i) = \frac{2E}{N} \tag{2.1}$$

The average degree quantifies globally what is measured locally by the vertex degree: the number of edges per node in G . N is usually referred to as the network size. Examples of values of N, E and $\langle k \rangle$ for some investigated biological networks are given in chapter 4, paragraph 4.1.4. For an integer $m \geq 1$, a *path* of

length m from a node v to a node u is a sequence of nodes $v = v_0, v_1, \ldots, v_m = u$ such that each pair (v_i, v_{i+1}) is in the edge set E. If m is the minimum length for all pathes from v to u then m is called the *minimal path length*. The *mean path length* (MPL in the table of paragraph 4.1.4) is defined as the average of the minimal path lengths for all pairs of nodes (v, u).

Moreover, each node v is characterized also by the clustering coefficient C_v. If v has at most one neighboor then $C_v = 0$. Otherwise C_v gives fraction of the triangles that go through node v over the total number of triangles that could pass through node v. It is equal to 1 for a node at the center of a fully interlinked cluster and it is 0 for a node that is part of a loosely connected group. Extending the concept, the average clustering coefficient of the graph (C in the table of paragraph 4.1.4) is defined as the average of C_v over all v; if its value is relatively high, this is a network property that is charachterictic for a modular organization. Therefore C is a measure of the network modularity.

Generally, graph types are classified by the *degree distribution* $P(k)$. Which is defined as the number of nodes with k links from the total number of nodes. We can so have exponential networks, major protagonist of this type being the random graph models, and scale-free networks. Recent studies have shown the relevance of scale-free connectivity for cellular network [57]. In particular, the protein protein interaction network generated by the two hybrid approaches of Ito and Huetz has also been found to have a scale-free topology [57]. The random graph models are homogenous networks with nodes comprising approximately the same number of links $k \sim \langle k \rangle$ [57]; they are characterized by P(k), which peaks at an average $\langle k \rangle$ and decay exponentially. By contrast the class of the scale-free networks is inhomogeneous and the connectivity distribution decay as a power-law $P(k) \sim k$. Compared with exponential networks the probability that a node is highly connected ($k >> \langle k \rangle$) is statistically significant in scale free networks [57].

2.2 Machine learning frameworks

Nowadays, one of the most challenging problems in computational biology is to transform the huge volume of data, provided by newly developed technologies, into knowledge. *Machine learning*, that has become an important tool to carry out this transformation, consists in programming computers to optimize a performance criterion by using example data or past experience. The optimized criterion can be the accuracy provided by a predictive model (in a modelling problem), the value of a fitness or an evaluation function (in an optimization problem).

2.2.1 Supervised classification

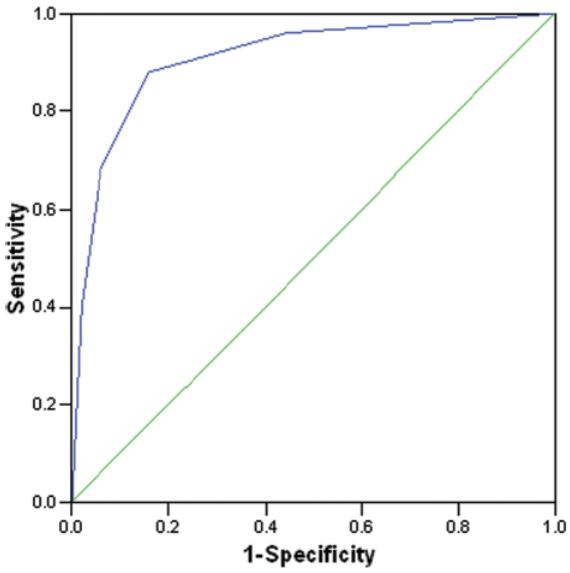
In a classification problem, we have a set of elements divided into classes. Given an element (or instance) of the set, a class is assigned according to some of the elements features and a set of classification rules. In many real-life situations, this set of rules is not known, and the only information available is a set of labelled examples (i.e. a set of instances associated with a class). Supervised classification paradigms are algorithms that induce the classification rules from the data. At this point, we can use the set of labelled examples as training set to build up a classifier. Once the classifier has been trained, we can use it to label new istances.

In two-group supervised classification, there is a feature vector XR^n whose components are called predictor variables and a label or class variable $C(0, 1)$. Hence, the task is to induce classifiers from training data, which consists of a set of N independent observations $D_N = (x^1, c^1)...(x^n, c^n)$ drawn from the joint probability distribution $p(x, c)$ as shown in figure 2.2.1. The classification model will be used to assign labels to new instances according to the value of its predictor variables.

	X_1	...	X_n	C
$(x^{(1)}; c^{(1)})$	$x_1^{(1)}$	...	$x_n^{(1)}$	$c^{(1)}$
$(x^{(2)}; c^{(2)})$	$x_1^{(2)}$	...	$x_n^{(2)}$	$c^{(2)}$
$(x^{(N)}; c^{(N)})$	$x_1^{(N)}$	...	$x_n^{(N)}$	$c^{(N)}$
$x^{(N+1)}$	$x_1^{(N+1)}$	...	$x_n^{(N+1)}$	???

When a 0/1 loss is used, all errors are equally bad, and our error calculations are based on the confusion matrix in figure 2.2.1. In this case, we can define the error rate as $(|FN|+|FP|)/N$, where $N = |TP|+|FP|+|TN|+|FN|$ is the total number of instances in the validation set. To tune a classifier, another approach is to draw the receiver operating characteristics (ROCs) curve [43], which shows hit rate versus false alarm rate, namely, $1 - specificity = |FP|/(|FP||TN|)$ versus $sensitivity = |TP|/(|TP||FN|)$, and has a form

		Predicted class	
		Positive	**Negative**
True class	Positive	TP: True positive	FN: False negative
	Negative	FP: False positive	TN: True negative



similar to Figure 2.2.1. For each classification algorithm, there is a parameter or a threshold of decision, which we can play with to change the number of true positives versus false positives. Increasing the number of true positives also increases the number of false alarms; decreasing the number of false alarms also decreases the number of hits. Depending on how good/costly these are for the particular application we have, we decide on a point on this curve. The area under the receiver operating characteristic curve is used as a performance measure for machine learning algorithms [44].

Optimization methods

Many problems in bioinformatics can be posed as the task of finding an optimal solution in a space of multiple (sometimes exponentially sized) possible

solutions: this is the context in which optimization methods find their natural collocation. Examples of approximate optimization algorithms or approximate optimization frameworks visiting one point of the search space at each iteration (*local search*) are: *Monte Carlo algorithm* [45],*Simulated annealing* [46] and *tabu search* [47] Examples of search methods using a set or population of points instead of a single point are: evolutionary algorithms (i.e *genetic algorithms, genetic programming, estimation of distribution algorithms or EDA* [48–50].

2.3 Information Theory to extract knowledge from multiple sequence alignments

Information and uncertainty are technical terms that describe any process that selects one or more objects from a set of objects. Suppose we have a device that can produce 3 symbols, A, B, or C. As we wait for the next symbol, we are *uncertain* as to which symbol it will produce. Once a symbol appears and we see it, our uncertainty *decreases*, and we remark that we have received some *information*. That is, information is a decrease in uncertainty. How should uncertainty be measured? The simplest way would be to say that we have an "uncertainty of 3 symbols". This would work well until we begin to watch a second device at the same time, which, let us imagine, produces symbols 1 and 2. The second device gives us an "uncertainty of 2 symbols". If we combine the devices into one device, there are six possibilities, A1, A2, B1, B2, C1, C2. This device has an "uncertainty of 6 symbols". If we take the logarithm of the number of possible symbols we can add the number of symbols instead of multiplying: the first device made us uncertain by $\log(3)$, the second by $\log(2)$ and the combined device by $\log(3) + \log(2) = \log(6)$. The base of the logarithm determines the units. When we use the base 2 the units are in bits (base 10 gives digits and the base of the natural logarithms, e, gives nats [51] or nits [52]). Thus if a device produces one symbol, we are uncertain by $\log_2 1 = 0$ bits, and we have no uncertainty about what the device will do next. If it produces two symbols our uncertainty would be $\log_2 2 = 1$ bit. In reading an mRNA, if the ribosome encounters any one of 4 equally likely bases, then the uncertainty is 2 bits. Summarizing, uncertainty is $\log_2(M)$, with M being the number of symbols. In case of symbols not equally likely, by rearranging the formula like this:

$$\begin{aligned} \log_2(M) &= -\log_2(M^{-1}) \\ &= -\log_2(\frac{1}{M}) \\ &= -\log_2(P) \end{aligned} \tag{2.2}$$

so that $P = 1/M$ is the probability that any symbol appears, we can deal with symbol appearing rarely relative to the other symbols. Generalizing this for various probabilities of the symbols, P_i, so that the probabilities sum to 1:

$$\sum_{i=1}^{M} P_i = 1. \tag{2.3}$$

The "surprise" that we get when we see the ith kind of symbol was called the "surprisal" by Tribus [53] and is defined by analogy with $-\log_2 P$ to be

$$u_i = -\log_2(P_i). \tag{2.4}$$

'u' stands for uncertainty; for example, if P_i approaches 0, then we will be *very surprised* to see the ith symbol (since it should almost never appear), and the formula says u_i approaches ∞. On the other hand, if P_i=1, then we won't be surprised at all to see the ith symbol (because it should always appear) and u_i = 0.

Uncertainty is the *average surprisal* for the infinite string of symbols produced by our device. To calculate he average surprisal for a string of length N that has an alphabet of M symbols we can follow this steps: Suppose that the ith type of symbol appears N_i times so that if we sum across the string and gather the symbols together, then that is the same as summing across the symbols:

$$N = \sum_{i=1}^{M} N_i. \tag{2.5}$$

There will be N_i cases where we have surprisal u_i. The average surprisal for the N symbols is:

$$\frac{\sum_{i=1}^{M} N_i u_i}{\sum_{i=1}^{M} N_i}. \tag{2.6}$$

By substituting N for the denominator and bringing it inside the upper sum, we obtain:

$$\sum_{i=1}^{M} \frac{N_i}{N} u_i \tag{2.7}$$

If we do this measure for an infinite string of symbols, then the frequency N_i/N becomes P_i, the probability of the ith symbol. Making this substitution, we see that our average surprisal (H) would be:

$$H = \sum_{i=1}^{M} P_i u_i. \tag{2.8}$$

Finally, by substituting for u_i, we get Shannon's famous general formula for uncertainty:

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i \tag{2.9}$$

In an MSA (multiple sequence alignment), the amino acids in a given column can be considered as a set of observations of a random variable x, with possible values defined by an alphabet K and associated probability distribution $p(X) = p(x_1), p(x_2), ..., p(x_K)$, where ,

$$\sum_{i=1}^{K} p(xi) = 1 \tag{2.10}$$

An estimate of the entropy $H(X)$ is obtained by using the observed amino acid frequencies, $f(xi)$, in place of the underlying probabilities, $p(xi)$ in equation 2.9. The concept of entropy can be easily extended to the case of two random variables where we have ordered pairs (x_i, y_j). In this instance, it is helpful to think of the pairs as elements of an extended alphabet, $K * L$, whose elements are all possible distinct pairs. If we have a pair of random variables, then joint or pair entropy is defined as follows:

$$H(X,Y) = \sum_{i=1}^{K} \sum_{j=1}^{L} p(x_i, y_j) \log_b p(x_i, y_j) \tag{2.11}$$

Conditional entropy is given by $H(X|Y) = H(X,Y)H(Y)$. Note that $H(X,Y) = H(Y,X)$, but that generally $H(X|Y)! = H(Y|X)$ (equality is obtained if and only if H(X) = H(Y)).

Mutual Information, $MI(X,Y)$, is the reduction of uncertainty (as measured by entropy) of random variable X given random variable Y

$MI(X,Y) = H(X) - H(X|Y)$

$= H(X) - H(X|Y)$

$= H(X) - H(X,Y) + H(Y)$

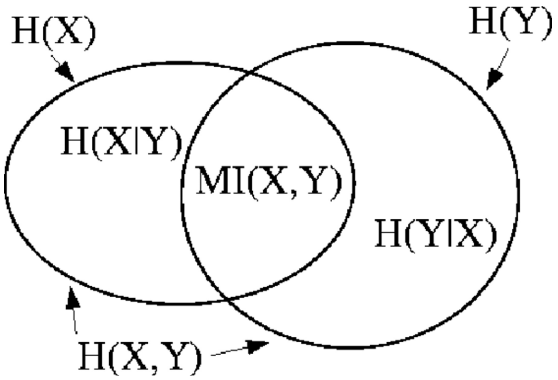
MI is symmetric, i.e. MI(X, Y) = MI(Y, X). Also note that H(X,X) = H(X), giving MI(X, X) = H(X).

Some important observations on H(X, Y) and MI(X, Y) may be proven but are most readily observed from the Venn representation of entropy space shown in figure 2.3.

By coming back to the multiple sequence alignments, the MI between two columns reflects the degree to which the pattern in the two columns is correlated. If amino acids occur independently at the two sites, the theoretical value for MI is zero. However, the estimate of the MI will only be zero if the observed pair frequencies reflects all possible pairings for the observed amino acid frequencies. Such a case is illustrated in Figure 2.3, where $X = A, C; f(A)\frac{1}{3}, f(C)\frac{2}{3}$ and $Y = E, M; f(E)\frac{1}{2}, f(M)\frac{1}{2}$. Adding or deleting a row would result in different pair frequencies and yield a non-zero value of $MI(X,Y)$, as would interchanging dissimilar amino acids within a column. When considering independent columns within finite MSAs, those alignments with fortuitous pairings resulting in vanishing MI values are the exception, as any deviation from the all possible

pairs" representation will yield a positive MI value. Also, as the number of amino acids occurring increases, the minimum sample size for which zero MI is possible increases quickly. If all 20 amino acids are present in each column and are equally probable, then zero MI is only possible if the frequency of each pair of amino acids is $1/202$. This condition cannot be met in an MSA with < 400 homologous proteins, and is unlikely to be met in columns of finite length.

Venn diagram of entropy space



Martin, L. C. et al. Bioinformatics 2005 21:4116-4124;
doi:10.1093/bioinformatics/bti671

Examples illustrating the effect of normalizing MI by pair entropy for MSAs with varied column entropies (area of circles) and dependencies (area of overlap)

	(a)	(b)	(c)	(d)	(e)
					
	AE	AE	AE	AE	AE
	AE	AE	AE	AE	AM
	AE	CM	CE	CE	CE
	CM	CM	CE	DM	CM
	CM	DF	DF	DF	CE
	CM	DF	DF	DF	CM
MI	.23	.37	.21	.23	0
$\frac{\text{MI}}{\text{H(X,Y)}}$	1	1	.58	.52	0

Martin, L. C. et al. Bioinformatics 2005 21:4116-4124;
doi:10.1093/bioinformatics/bti671

Chapter 3

Proposed frameworks for the proteome wiring problem

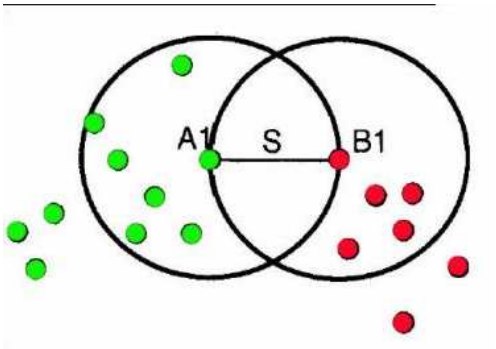
In this and in the following chapter, we describe some computational frameworks for the wiring of proteomes that is the subject of this thesis. In particular, this work is aimed to developing computational frameworks to calculate the probabilities p_ij, weights of the edges linking all possible combinations of protein pairs in human and yeast protein graphs, according to the definitions given in chapter 2. For each examined approach, following a general description, there are some paragraphs describing the material and methods used to obtained the results illustrated in the chapter 4.

3.1 Human proteome wiring by using interologs in model organisms: the HomoMINT interactome

Predictions using interologs [71], are based on the theory that proteins interacting in one organism co-evolve such that their respective orthologs mantain the ability to interact in another organism. This concept was originally proposed by Walhout [81]. Matthews et al. [82] used the same approach to map yeast interactions into C.elegans proteome and then verified them by yeast two-hybrid.

In chapter 4 we report in some details how the interolog concept has been applied in building the predicted human interactome HomoMINT [70].

In the following paragraphs are reported the material and methods employed for this work.



Schematic illustration of the Inparanoid algorithm used in this work to estabilish the ortholog relationships between interacting proteins. Each circle represents a sequence from species A (green) or species B (red). Main orthologs (pairs with mutually best hit) are denoted A1 and B1. Their similarity score is shown as S. The score should be thought of as reverse distance between A1 and B1, higher score corresponding to shorter distance. The main assumption for clustering of in-paralogs is that the main ortholog is more similar to in-paralogs from the same species than to any sequence from other species. On this graph it means that all in-paralogs with score S or better to the main ortholog are inside the circle with diameter S that is drawn around the main ortholog. Sequences outside the circle are classified as out-paralogs. In-paralogs from both species A and B are clustered independently.

3.1.1 Data Sources

The proteome sets for the BLAST searches and ortholog table assembling were downloaded or built from the following web sources: Arabidopsis thaliana proteome set (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=3
Caenorhabditis elegans (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=9
Drosophila melanogaster (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=17
Escherichia coli K12 (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=18
Homo sapiens (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=25
Mus musculus (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=59

Rattus norvegicus (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=122
Saccharomyces cerevisiae (predicted proteins),
http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeID=40
Multiple species proteome set (predicted proteins),
http://mint.bio.uniroma2.it/mint/ by querying the database for proteins belonging to the following species: Sus scrofa (Pig), Xenopus laevis (African clawed frog), Ovis aries (Sheep), Oryctolagus cuniculus (Rabbit), Gallus gallus (Chicken), Canis familiaris (Dog), Bos taurus (Bovine).

3.1.2 Software

BLASTP searches were carried out using blastall 2.2.9 [63]. InParanoid algorithm version 1.35 was downloaded from: http://inparanoid.cgb.ki.se/index.html. Graph analysis and GO functional annotation analysis were performed by using R package version 2.0.1 from http://www.r-project.org/ and the Bioconductor modules graph, RBGL, GOstats [65].

3.1.3 Assembly of the orthology table

The procedure implemented in the InParanoid algorithm [73] starts with an all-against-all BLASTP comparison between two proteomes of interest. Reciprocal best hit criteria are used to identify orthologous relationships between pairs of proteins. For each putative ortholog, probable recent paralogs or in-paralogs are identified as sequences within the same proteome that are reciprocally more similar to each other than either is to any sequence from the other proteome. An InParanoid confidence level cut-off of 0.6 was chosen for the assignment of in-paralogs to orthology groups. Due to the redundancy of the starting proteome sets, several groups contained identical copies of the same protein. To limit this problem we decided to eliminate paralogs with InParanoid confidence level above 0.98. InParanoid performs its comparison between each pair of proteomes. To build an orthology table with orthology groups including proteins from all organisms of interest, we used python scripts to merge the InParanoid results keeping a human protein as reference for each orthology group.

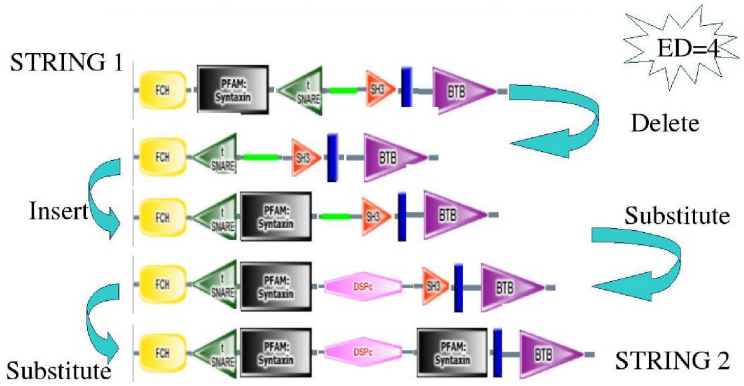
3.1.4 Assembling HEN (Human Experimental Network)

The human experimental interactome has been assembled by importing the data in a Postgresql database from the following resources: Intact (XML PSI files), 1300 unique interactions at http://www.ebi.ac.uk/intact/index.jsp, DIP (Flat file), 833 unique interactions at http://dip.doe-mbi.ucla.edu/, BIND (XML PSI 2 file), 4073 unique interactions at http://bind.ca/, MINT, 3679 unique interactions at http://mint.bio.uniroma2.it/mint/, HPRD (XML PSI file), 6153

unique interactions at http://www.hprd.org/, MIPS (XML PSI file), 322 unique interactions at http://mips.gsf.de/proj/ppi/. Only interactions that could be confidently mapped to Uniprot identifiers were added to HEN.

3.1.5 Filtering orthology groups for domain architecture homogeneity

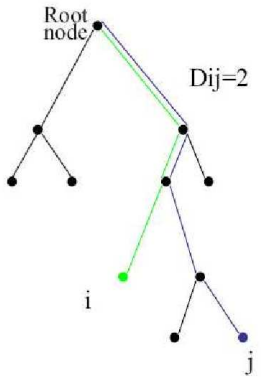
A procedure has been developed to improve and to measure the functional coherence in orthology groups, based on dynamic programming techniques and implemented as a string matching algorithm [66]. We modeled every protein in our orthology groups as an ordered string of domains. To this end, we used the domain annotations available in SMART [67] and PFAM [68]. In particular, the human and the other eight model organism proteomes under analysis have been surveyed for their specific domain architectures. Repetitions of the same domain are treated as a single instance of that domain. Overlapping domains are considered as independent elements of the string representing the domain architecture of the protein. Then we developed a PERL string matching algorithm to establish distances between the proteins in terms of similarities between their domain architectures. Each protein is represented as a string of concatenated ordered domains. Thus we were able to measure a distance between two proteins by counting the number of domain editing steps (deletions, insertions, substitutions) in order to match the domain architecture of the two proteins. Proteins identical in their domain architecture will have an "edit distance" equal to zero. Distances are normalized by dividing for the total number of domains in the ortholog human protein. This procedure prevents proteins with markedly different domain architecture (and function) from being clustered mistakenly in a group, although they share similarities only within distinct regions of a multidomain protein. In this way we tried to take in account not only local relationships among sequences to be merged in the orthology groups but global relationships as well. To assess the filtering procedure we examined the consistency of the annotation of the members within each orthology group, as reported in the ENZYME database [69]. We were able to attribute at least two ENZYME annotations to 9% of groups constituting the filtered orthology table. Fewer than 6% of these groups (77/1355) were declared inconsistent with the ENZYME hierarchic classification scheme. 17 inconsistent groups present in the standard orthology table were not present in the filtered orthology table, underlining the improvement of the functional coherence in the orthology groups after filtering for similarity in domain architectures. The number of inconsistent groups in the standard orthology table was 94 out of 1396 groups which have at least two ENZYME annotations.



Schematic representation of the algorithm used to assess the homogeneity in the domain architecture in the ortholog groups;the edit distance between the two strings or proteins is the number of substitutions insertions deletions to go from string 1 (protein 1) to string 2 (protein 2).

3.1.6 Gene Ontology similarity analysis

The algorithm for measuring the Gene Ontology annotation similarity of a pair of proteins is based on the simLL function of the GOstats package of Bioconductor [65]. For each pair of proteins (Pi, Pj) and for each ontology, the function simLL assigns, in three steps, a unique measure of similarity, called Dij (depicted in Fig. 4.1.4): (1) Finds all the terms to which Pi and Pj are annotated including the parent terms. These sets of terms in the Gene Ontology tree represent the nodes of the GO graphs induced by Pi and Pj, respectively. (2) Find the set of terms which the GO graphs induced by Pi and Pj have in common. Denote this set Sij. (3) Define the depth of each term in Sij to be the length of the shortest path between the term and the root node of the ontology (here length refers to number of connecting edges). (4) Find the maximum depth of terms in the set Sij. We refer to this value as Dij.



Schematic representation of the algorithm used to evaluate the relatedness of gene ontology annotation. The Gene Ontology graph induced by protein i is in green, while the one induced by protein j is in blue. Dij is the number of edges that the two induced graphs have in common.(image from [70]).

3.2 Wiring of proteome in yeast by data integration

In yeast (*S. cerevisiae*), attempts to solve the proteome wiring problem have involved a lot of experimental and computational efforts. Many questions remain still unanswered in spite of the abundance of protein interaction data from high-throughput experiments: how to wire those portions of the yeast interactome for which the available experimental data show embarrassing incongruencies in their results and coverage? How to deal with the different kind of experimental error associated with the different experiments and techniques? Moreover, leaving the experimental point of view and moving towards a more theoretical point of view, is it possible to define an a priori for protein interaction? i.e. is it possible to estimate the probability that two proteins selected at random from a proteome are found functionally linked or connected in our protein graph? In addition to experimentally determined interaction data sets, there exists a large amount of biological information in the form of sequence, structure, functional annotation, expression level and functional genomics data sets. The second part of this thesis is dedicated to exploit the interaction evidence contained in some functional genomics data sets with the aim to obtain a system biology

a priori for protein interaction to be included in a more complex model integrating in a single probabilistic framework the a priori for protein interaction and an estimation of the specific experimental error and coverage inherent to the different experimental approaches used to detect protein interaction in the different available data sets. The final purpose of this work is to end up with a flexible and simple software available to everybody in the scientific community is interested in assembling on own computer and with some well defined criteria the huge amount of experimental protein interaction data sets [**?**]. Following this section, materials and methods to obtain the results presented in chapter 4 are described.

3.2.1 Data sources

Data sources: functional genomics data sets

- Protein concentration data:
  Recently two large studies have addressed the problem of protein concentration in a yeast cell at a proteomic level (6235 tagged proteins). These studies [94,95], have identified 4517 expressed proteins. For 3868 of these, the concentration could be quantified by TAP western. Some detected proteins were not quantified because they were either only seen by GFP fusion (66) or because of very low expression levels (234) , or because of experimental problems (149). 525 of the proteins that could not be detected were classified as spurious, on the basis of their codon usage while the remaining 169 remained as "undetected but possibly bona fide" proteins. We have assigned the minimal concentration value of 40 proteins/cell to the 234 proteins that were detected but whose concentration could not be measured because of low levels. The missing experimental values for a total of 384 proteins of unknown concentration (the remaining 66 GFP fusion, 149 "Experimenal problems" plus 169 "undetected but annotated" were artificially added to the data set by using the "impute" function in the Hmisc R package.

- mRNA concentration data:
  This dataset represents the time course of expression fluctuations during the yeast cell cycle and Rosetta compendium, consisting of the expression profiles of 300 deletion mutants and cells under different chemical treatments [80].

3.2.2 Software

Graph analysis and GO functional annotation analysis were performed by using R package version 2.5.0 and the Bioconductor modules graph, RBGL, GOstats

[65]. The "'descriptive statistical analysis" of the predictors modelled from the functional genomics data sets were performed by using the statistical package R (http://www.r-project.org). The logistic regression model was built by using the R packages Hmish and Design (http://www.r-project.org). The model evaluation was performed with the R package ROCR (http://www.r-project.org).

3.2.3 Assembling the gold standard

Positives examples comes from a recent completion of a comprehensive literature curation effort made available at http://wwww.biogrid.org, [91]. In particular, the data set named LC_BIOCHEMICAL_011906_MinusRNA.txt was used: it contains the Literature Curated biochemical interactions minus a few RNA genes; the BIOCHEMICAL systems included are:
Affinity Capture MS
Affinity Capture Western
Two-hybrid
Colocalization
FRET
Reconstituted Complex
Protein-peptide
Co-purification
Co-fractionation
Biochemical Activity
Co-crystal Structure
Far Western
Protein-RNA
The authors specified that all co-purification complexes are strictly represented in the datasets as a minimal spoke model. Because bait assignment for co-purification complexes is arbitrary (i.e. there is no bait per se), the hub of each spoke was defined by the most highly connected protein within each complex defined by the LC_BIOCHEMICAL set minus the Co-purification information; if no information was found in the LC_BIOCHEMICAL set, then the HTP_BIOCHEMICAL set was used. All the datasets available from http://wwww.biogrid.org, are formatted as tab delimited text files with the following header:
Bait gene protein , Hit gene protein, Bait Standard Name, Hit Standard Name, Experimental System , Source, PubMed ID

According to the suggestion of [96], negatives examples were generated by using the algorithm randomEgraph in the package graph of the Bioconductor suite [65].

3.3 Identification of protein interactions in human by co-evolutionary information of the interacting partners

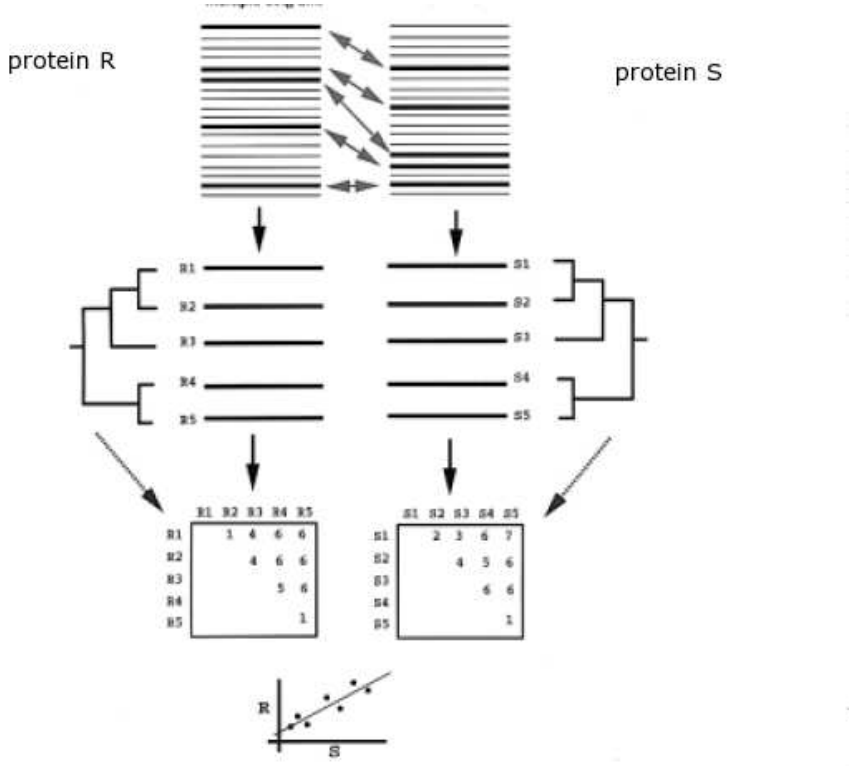
The prediction of protein-protein interaction with genomic information is an important issue of bioinformatics. A protein sequence has an information content and it is possible to unravelling it by studying its evolution during the time. The multiple sequence alignment of orthologues is a very interesting object to study information content of a protein (see chapter 2). Studies of protein evolution have mainly focused on unravelling the selective constraints acting on protein-coding genes, more recently, on genes coding for interacting proteins. Because the interacting proteins tend to co-evolve as result of the selection pressure exerted by the interaction, if we are able to unravel the co-evolutionary signals, we are also able to predict interaction between two proteins by modeling and representing these signals in an appropriate way.
The set of recently developed bioinformatics methods that utilize co-evolutionary information of the interacting partners can be divided in two broad approaches:

1. site-specific methods [107, 108]

2. methods based on the analysis of the entire amino acid sequence. [98]

Methods belonging to the first category are those based on Information Theory and information content detection in multiple sequence alignments (see chapter 2).

A description of the most used method belonging to the second category (the mirror tree) is given in fig. 3.3.



Scheme of the mirror tree method [98]. The initial multiple sequence alignments of the two proteins are reduced, leaving only sequences of the same species and consequently the trees constructed from these reduced alignments would have the same number of leaves and the same species in the leaves. From the reduced alignments, the matrices containing the average homology for every possible pair of proteins are constructed. Such matrices contain the structure of the phylogenetic tree. Finally, the similarity between the data sets of the two matrices and implicitly the similarity between the two trees are evaluated with a linear correlation coefficient.

3.3.1 Data Sources

The orthology sets for the co-evolutionary analysis were downloaded and built by using the data from the web resource Ensembl at http://www.ensembl.org and a set of APIs written in perl very usefull in querying and retrieving the data from the Compara section of the Ensembl resource.
The protein interaction data for the assembling the *Human SH2ome* were downloaded in form of flat files from the following resources: BIOGRID, HPRD, DOMINO (MINT), Intact [28,30,31,33]. More in details: Intact(Flat file)contributed with 1300 unique interactions, DOMINO (MINT) with 3679 unique interactions,

HPRD (Flat file) with 6153 unique interactions, BIOGRID(Flat file) with 322 unique interactions. Only interactions that could be confidently mapped to Ensembl ids were added to the *SH2ome*.

3.3.2 Software

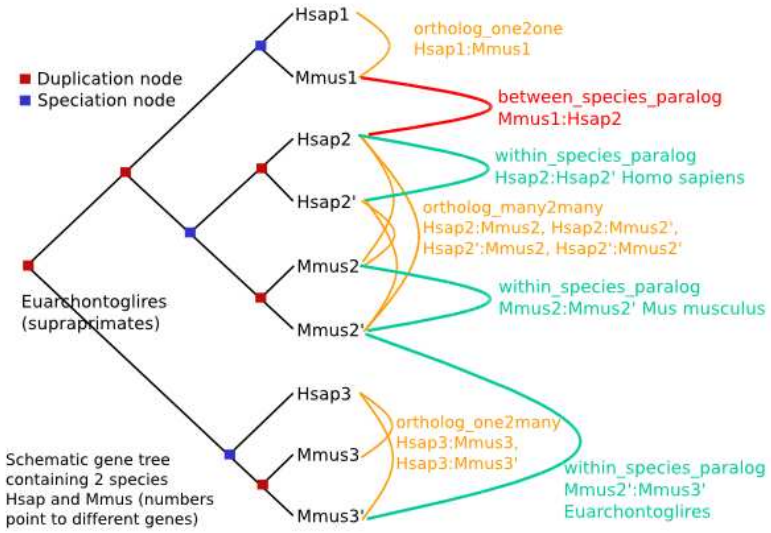
The igraph suite for graph analysis was downloaded from:
http://cneurocvs.rmki.kfki.hu/igraph/.
Descriptive statistical analysis were performed by using R package version 2.5.0 from http://www.r-project.org/.
Protein Evolution Covariation Analysis was performed by PECA a highly modularized code written by Mario Fares and Francisco Codoner; it runs on Linux/Unix, Windows and Mac Os X operating systems; functions are very well split into different modules for an easier addition of new functions.
Correlated evolution analysis by the mirror tree approach was performed by using a modified version of the pipeline developed by Jothy [103].

3.3.3 Assembly of the *SH2ome*, a set of human interacting proteins in which at least one in every interacting protein pair is an *SH2 domain containig* protein

The human experimental *SH2ome* has been assembled by importing the data in form of flat files from the following resources: Intact(Flat file),1300 unique interactions at http://www.ebi.ac.uk/intact/index.jsp, DOMINO, 3679 unique interactions at http://mint.bio.uniroma2.it/mint/ HPRD (Flat file) , 6153 unique interactions at http://www.hprd.org/ BIOGRID(Flat file), 322 unique interactions at http://mips.gsf.de/proj/ppi/ Only interactions that could be confidently mapped to Ensembl ids were added to *SH2ome*.

3.3.4 Criteria for the assembling of the orthology sets

The followed procedure was that implemented in the Compara pipeline. [59].
The gene orthology and paralogy predictions are generated by a pipeline where maximum likelihood phylogenetic gene trees (generated by TreeBeST) play a central role. The trees aim to represent the evolutionary history of gene families, i.e. genes that diverged from a common ancestor. They reconciled with their species tree have internal nodes annotated to distinguish duplication or speciation events. There is a clear concordance with reciprocal best approaches in the simple case of unique orthologous genes. However, the gene tree pipeline is able to find more complex one-to-many and many-to-many relations.



Examples on more complex than one-to-one relationship: one-to-many and many-to-many relations.The pipeline used to obtain the ortholog sets is able to find these complex sets.

Chapter 4

Results and discussion

In this chapter the complete set of results obtained by using different approaches to the problem of wiring proteomes will be presented and discussed. In the next chapter we will deal with final considerations and future perspectives.

4.1 Results and discussion for the reconstruction of human protein interolog network using evolutionary conserved networks: the HomoMINT interactome

We first outline results and discussion relative to the assembling of the HomoMINT interactome, reported in detail in [70].

Our strategy, starts by assigning model organism proteins to orthology groups having a human protein as the main ortholog. An interaction between human proteins is then inferred if both partners of an interaction experimentally verified in model organisms have at least one human ortholog. Similarly to Lehner and Fraser [72], we have used the InParanoid algorithm [73] to assemble orthology groups. This algorithm has the potential to distinguish between outparalog, homologous genes that arose by duplication before the speciation event (unlikely to share function), and in-paralogs arising after speciation. However, to avoid unnecessary graphical overcrowding in the visualization of the inferred network, we have only included in the resulting predicted human network (HomoMINT) the interactions between the main human orthologs of each orthology group. An extended network in which the model organism interactions are mapped to all the possible combinations of in-paralogs is also available (see Additional Material in [70]). Since InParanoid attributes a score to each orthology

assignment it is relatively easy to obtain different inferred networks using orthology tables with varying levels of stringency for assignment to orthology classes. In addition we have tuned the orthology assignments by imposing the condition that proteins in the same orthology group must have the same domain architecture. This filtering step evaluates the overall protein similarity and eliminates any incongruity caused by the local nature of the BLAST algorithm. Motivated by the observation that multidomain proteins, sharing an exact domain architecture, have significantly higher functional conservation [74, 75] , we developed a workflow (see paragraph 3.1.5) to produce a "high confidence" orthology table in which all orthology group members share the same domain architecture. This filtering procedure improves the functional coherence within the orthology groups, while removing only 10% of the 16531 inferred groups. We call the resulting network HomoMINT_filtered(see Additional Material in [70]).

4.1.1 Intersection of HomoMINT with the Human experimental network

Several low throughput experiments, providing evidence of protein interactions between human proteins, have been published in the scientific literature over the past decades. This dataset is approximately the same size as the datasets obtained from the results of high throughput experiments carried out in model organisms, although it is not readily accessible. Recently, a number of databases have started to capture this information and release it in a computer readable format according to a common standard [70]. By merging all the interactions currently deposited in seven major databases [70], we have assembled a human interactome of 28531 non-redundant interactions. In Table 4.1 we have reported the analysis of the overlap between the data curated by the different databases. This assembled human experimental network (HEN) is likely to have some bias in the coverage of the interaction space due to the interest of the scientific community in investigating specific biological domains or to a biased selection of the journal articles curated by the databases. Nevertheless it represents the most accurate representation of the human interactome to date. We used HEN as a benchmark for the initial assessment of the accuracy and the information content of HomoMINT and related inferred networks (Table 2). The networks inferred by Brown and colleagues [70] and by Lehner and Fraser [70] are here referred to as "OPHID" and "Sanger" respectively. As proposed by Marcotte and colleagues [70] we used a unified scoring scheme to evaluate the ability of each inferred network to reconstruct the reference network. To evaluate a dataset we calculated a log likelihood ratio as where $P(I|D)$ and $P(\ I|D)$ are the frequencies of interactions, in a given dataset (D), that are or are not observed in the benchmark dataset (I), while $P(I)$ and $P(\ I)$ represent the prior expectations (the

frequency of all benchmark gene pairs that do or do not interact). The overlap between the human experimental network and the one inferred from model organisms (HomoMINT) is 694 interactions (Table 3). This corresponds to 7.1% of HomoMINT, suggesting that both networks only cover a small fraction of the real interactome and that either or both are affected by a large number of false positives. Most of the HomoMINT network (94%) is inferred from interactions that have been obtained by high throughput experiments while only 6% is inferred from higher confidence experiments. Interestingly, the set of high confidence interactions covers more than 26% of the intersection between HomoMINT and the experimental network. The OPHID and Sanger networks are larger since their inference is based on a larger dataset, including computationally predicted interactions datasets (Sanger), and binary interactions, within complexes, being represented by the matrix [70] model (OPHID). This results in a much larger number of binary interactions than for instance those present in networks based on the 'spoke' model. As a consequence the coverage of the HEN network is also larger but the percentage of confirmed interactions and the LLR is lower when compared with HomoMINT. The Sanger core dataset, whose inference is based on a subset of high confidence interactions, is more accurate as is the HomoMINT high confidence network containing only interactions that are inferred when supported by at least two experiments. The highest log likelihood ratio is achieved by a rather limited network HMINT_2org (126 edges) where we have only considered the interactions confirmed by experiments in at least two model organisms. The overlap between the human experimental network and HomoMINT_filtered, obtained by considering only ortholog pairs sharing the same domain architecture, is 453 interactions; these corresponding to almost 9% of the inferred interactions.

		MINT	**DIP**	**BIND**	**Intact**	**React.**	**HPRD**	**MIPS**
	Nr. of edges							
MINT	3679	x	315	340	1350	101	429	54
DIP	990		x	158	22	67	20	26
BIND	4671			x	356	229	733	50
Intact	2860				x	103	208	16
React.	15068					x	269	16
HPRD	6891						x	84
MIPS	777							x

Table 4.1: analysis of the overlap between the data curated by the different databases.

4.1.2 Intersection of HomoMINT with the iHOP resource

The PubMed resource, containing more than 15 million biomedical abstracts, is a valuable resource for high quality protein interactions. As a whole, concurring proteins in PubMed sentences can be considered and modeled as a lit-

Dataset	Number of interactions	Description or reference
OPHID	23359	[77]
Sanger	37007	[72]
Sanger H.C.	5647	[72]
HomoMINT	9749	[70]
HomoMINT_filtered	5203	*
HMINT_2 int	290	**
HMINT_2 org	126	***
HM_LT	543	****
HEN	28531	*****
iHOP	278452	[76]

Table 4.2: Inferred ad experimental networks compared in this study; * HomoMINT filtered for domain architecture conservation , ** inferred from interactions confirmed by at least two experiments. , *** inferred from interactions supported by experiments in at least two model organisms , **** Inferred from interactions discovered by low throughput experiments. , ***** Compilation of interactions between human proteins

		OPHID	**Sanger**	**HEN**	**%overlap**	**LLR**
		23359	37007	28531		
HomoMINT	9749	3501	2794	694	7.1	4.2
OPHID	23359		7067	1632	7.0	4.1
Sanger	37007			1504	4.1	3.6
Sanger H.C.	5647			841	14.9	5.0
HM_filtered	5203	1818	1391	453	8.7	4.4
HMINT_2int	810	290	227	218	26.9	5.7
HMINT_2org	126	70	75	60	47.6	6.6
HM_LT	543	69	63	131	24.1	5.6

Table 4.3: Overlap between inferred and experimental human networks. For this comparison we mapped all the proteins to Uniprot ids. In this process proteins (and their interactions) that could not be confidently mapped were eliminated from the networks.

erature network, which can be superimposed on experimental interaction data or on putative relationships, making it possible to compare new and existing knowledge possible. Here we have made use of a novel text-mining resource, called iHOP (Information Hyperlinked over Proteins) [70] as an independent assessment of the protein interactions predicted in HomoMINT. The iHOP system currently contains 6 million sentences from PubMed abstracts and about 40000 different proteins from human, mouse, and other common animal models (iHOP, http://www.pdg.cnb.uam.es/UniPub/iHOP/). Table 4 summarizes the results obtained from this comparison. In particular, we were able to identify a corresponding sentence in the iHOP network for 6.8 % of our predicted interactions. Moreover, 3 % of these sentences expressed the interaction in an explicit protein-verb-protein syntax. In the control set (H_MINT ctrl), derived from a process of scrambling of the true dataset, less than 1 % of the putative interactor pairs were supported by co-occurrence in sentences in the iHOP database. For comparison the overlap of the iHOP human protein interaction network with

our assembled experimental PPI dataset was estimated to be about 22%. Only sentences of high precision were used for the assessment; sentences were excluded from the comparison, when ambiguities between protein-synonyms from different organisms (e.g. Mtx2 in mouse and MTX2 in human) could not be resolved. For this comparisons we mapped all the proteins to Locus Link ids. In this process proteins (and their interactions) that could not be confidently mapped were eliminated from the networks. For this reason, H_MINT in Table 4 contains 7658 interactions.

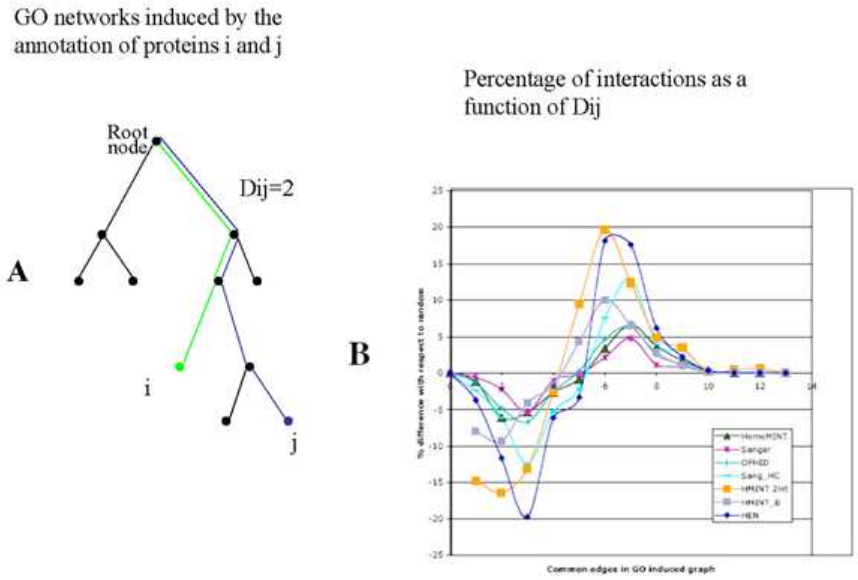
		H_MINT	**H_MINT ctrl**	**Sanger**	**Sanger ctrl**
	Nr. of edges	7658	7658	26590	26590
iHOP sentence	278452	522 (6.8)	57 (0.7)	857 (3.2)	233 (0.8)
iHOP pattern	47807	229 (3)	9 (0.1)	254 (1)	53 (0.2)
		OPHID	**OPHID ctrl**	**HEN**	**HEN ctrl**
	Nr. of edges	12887	12887	23332	23332
iHOP sentence	278452	941 (7.3)	88 (0.7)	5293 (22.6)	615 (2.7)
iHOP pattern	47807	468 (3.6)	14 (0.1)	2675 (11.5)	176 (0.7)

Table 4.4: The iHOP sentence network includes interactions between proteins whose names are found in the same sentence in an abstract. iHOP pattern is a subnetwork linking proteins found in a pattern of type gene_name_A verb gene_name_B. The networks that are compared with iHOP are described in the main text. The corresponding ctrl networks are scrambled networks containing the same nodes and the same number of edges. For this comparison we mapped all the proteins to Locus Link ids. In this process proteins and their interactions that could not be confidently mapped were eliminated from the networks.

4.1.3 Interacting proteins sharing GO terms

The extent of shared annotation in a protein interaction dataset has been previously shown to correlate with accuracy [**?**]. Thus, as a third benchmark for the assessment of the different inferred networks, we estimated the similarity of the Gene Ontology annotation (Biological Process) [?] of any pair of interacting proteins. To determine the relatedness of two GO terms we used the simLL function of the GOstats Package of Bioconductor [?]. This algorithm, as schematically illustrated in Figure **??** A, ecompares the GO graphs 'induced' by two proteins (i, j) and counts the number of edges that are in common between the minimal paths linking the two GO annotation nodes and the ontology root nodes. This value, Dij, is taken as a measure of annotation relatedness. Figure **??** B reports, as a function of Dij, the difference between the percentage of interaction pairs showing a given level of GO annotation similarity in an inferred network and in a comparable randomized network. In the randomized network the interactions between the same nodes were reassigned at random. All the inferred networks show a significant difference as compared to the scrambled networks, with the function peaking at Dij = 6 or 7. As was observed in the previous assessment

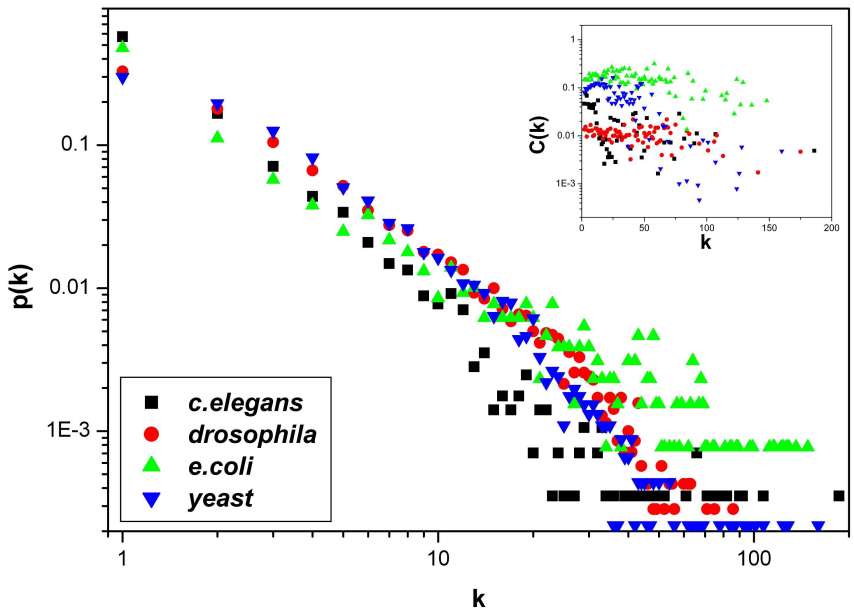
tests, the HomoMINT and OPHID networks perform better than the Sanger dataset, while the Sanger high confidence curve is more similar to the curve of the experimental network. A higher peak at Dij = 7 is observed in the curve of HomoMINT_filtered, obtained by filtering the orthology groups to remove proteins displaying a different protein architecture, or in the curve of HMINT_2int, a high confidence network obtained by considering only interactions supported by at least two experiments.



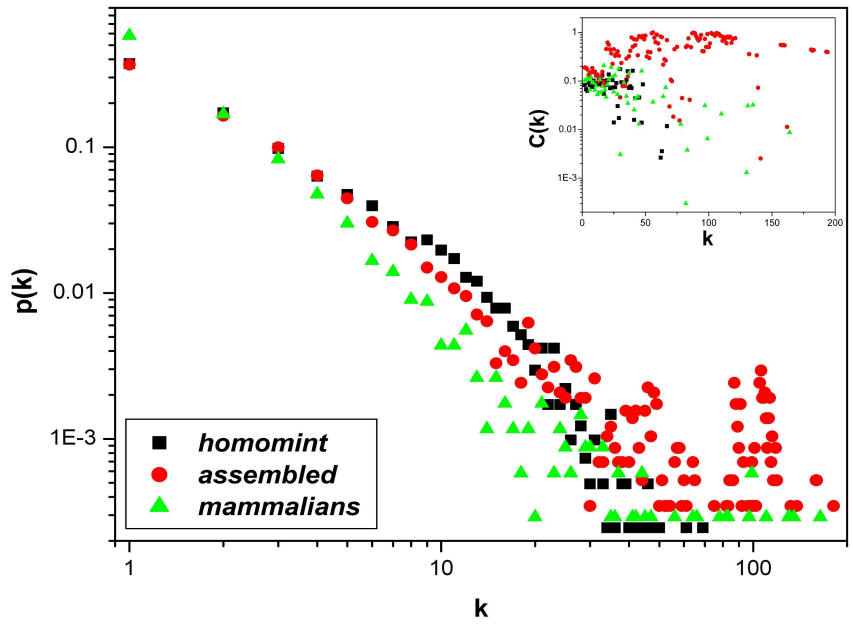
Degree of common annotation in interacting protein pairs in experimental and inferred networks. A. Schematic representation of the algorithm used to evaluate the relatedness of gene ontology annotation. The Gene Ontology graph induced by protein í ís in green, while the one induced by protein j́ís in blue. Dij is the number of edges that the two induced graphs have in common. B. For any given network we have derived a 'scrambled network' containing the same protein nodes linked by the same number of edges with their connections rearranged at random. For each interacting protein pair, in which both proteins have a GO annotation, we have then calculated Dij. Finally we have plotted, as a function of Dij, the difference between the percentage of nodes having a specific Dij in the inferred and in the scrambled network.(image from [70]).

4.1.4 HomoMINT as a graph

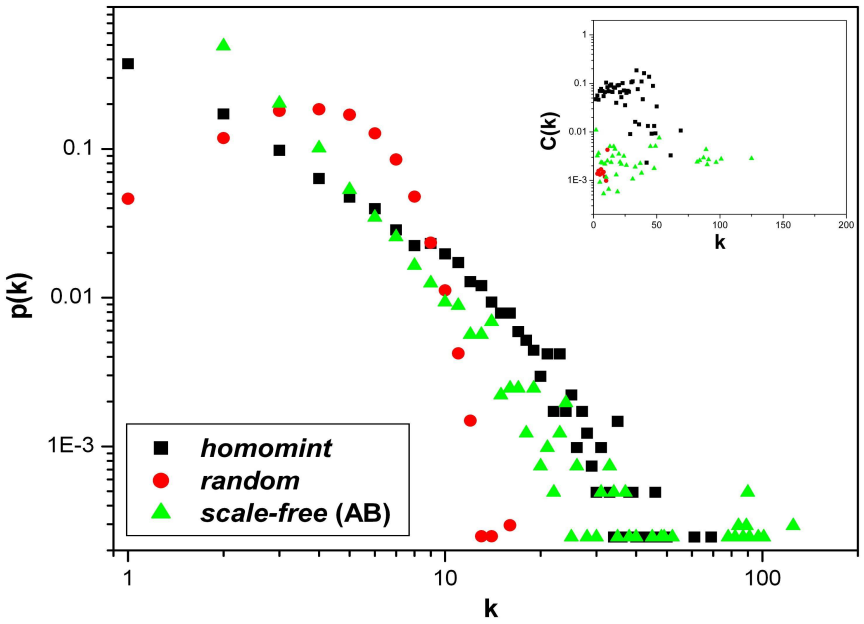
Protein interaction networks can be described as graphs where nodes and edges represent proteins and their interactions respectively. Although, at a first sight, apparently random in their topology, biological networks are characterized by a number of properties differentiating them from random networks. Specifically they have a large average clustering coefficient [57]. Most remarkably the distribution of protein connectivity is scale-free. As shown in Figure 4.1.4 the HomoMINT network, as well as the assembled human interaction network, has a scale-free topology with its degree distribution not differing substantially from those of the interactomes of model organisms. In Table 4.1.4 we have reported the analysis of some characteristics of the HomoMINT graph and we have compared them with those of some experimental networks in the MINT database. In HomoMINT the average clustering coefficient, the parameter that most captures the modularity of biological networks, is considerably higher than that of a random network of similar size and is consistent with the values found in biological networks. Also the remaining parameters describing the HomoMINT graph are typical of biological networks.



Degree distribition of the HomoMINT network compared with different biological networks.



Degree distribition of the HomoMINT network compared with mammalian network.



Degree distribition of the HomoMINT network compared with random network.

Data set	Nr prot (N)	Nr interactions (L)	$C(k)$
HomoMINT	3423	8266	0.05
HomoMINT_HC	2749	5631	0.05
HomootherDB	4674	10769	0.08
MamMINT	3445	5105	0.04
CaEMINT	2834	4406	0.02
DroMINT	7005	20282	0.01
YeastMINT	4584	12055	0.07
EcoliMINT	1289	5420	0.08
EpyloriMINT	698	1348	0.01
Random2000	1989	5047	0.002
Random5000	4893	9935	0.001

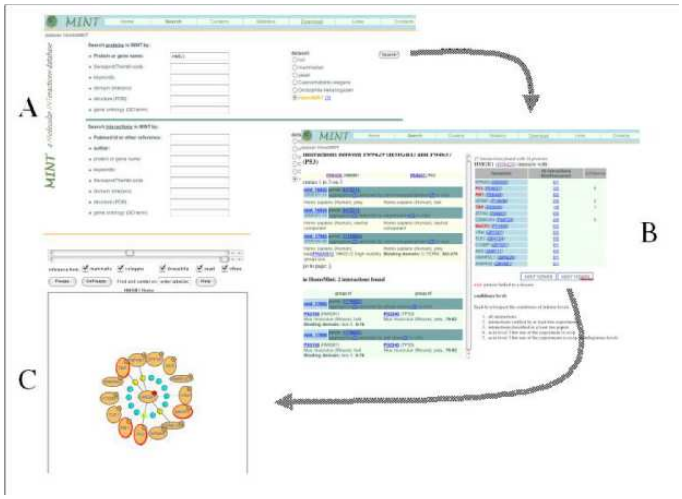
The function $C(k)$, the average clustering coefficient of all nodes with k links for the experimental networks has a behaviour completely different from the behaviour of the same function for random networks of similar size.

Data set	n conn_comp	N_LCC	L_LCC	d_LCC	MPL	$\langle k \rangle$
HomoMINT	91	3215	8148	14	4.811192	5.06874
HomoMINT_HC	96	2531	5508	15	5.048073	4.35243
HomootherDB	113	4408	10609	15	4.809685	4.813521
MamMINT	310	1486	1904	19	7.452652	2.562584
CaEMINT	100	2589	4265	13	4.850238	3.294708
DroMINT	62	6860	20135	11	4.46128	5.870262
YeastMINT	54	4449	11869	12	4.45502	5.335581
EcoliMINT	11	1255	5396	8	3.598813	8.599203
EpyloriMINT	8	684	1341	9	4.139075	3.921053
Random2000	1	1989	5047	11	4.865727	5.074912
Random5000	3	4888	9932	13	6.283034	4.063830

The prefix Mam stands for mammalian, Dro stands for Drosophila melanogaster, CaE stands for Caenorabdits elegans .

4.1.5 HomoMINT as a web server

The inferred HomoMINT network has been incorporated into the MINT database [88]. In essence HomoMINT is a calculated table integrated in the MINT relational database. The table is calculated every day by using the orthology group table to map onto the human proteome the interactions that are curated daily in the MINT database. As a result HomoMINT is a dynamic dataset continuously updated that can make use of the search and analysis tools developed for MINT. By entering a protein name, in the MINT search form, one can either perform the search over the experimentally verified interactions between human proteins, as curated in the MINT database, or extend the search to the HomoMINT table, by checking the appropriate radio button. In the latter case one obtains, as a result of the query, both the experimentally verified interactions and the inferred ones. Appropriate links make it possible to retrieve information about the experiments supporting the interaction either directly (experiments carried out with human proteins) or indirectly (experiments carried out in model organisms) (Fig. 1B). During any MINT search session it is possible to extend the analysis to HomoMINT, by clicking the HomoMINT hyperlink. The composition of the orthology groups used to infer the human interactions can also be inspected via the 'orthology table' hyperlink. A distinction is made between main orthologs (orthologs) and co-orthologs (in-paralogs). Finally the HomoMINT network can be analyzed, expanded, edited in the context of the experimentally verified protein interactions in the MINT database by using the MINT viewer tool (Fig. 1c). For instance the MINT viewer makes it possible, by checking appropriate boxes, to visualize only interactions inferred from any combination of model organism interactomes. The network visualized and edited by the viewer tool can be downloaded in any of three formats: flat file, XML PSI [70], or in a format that can be used as input for the OSPREY visualization software [70].



HomoMINT as a web tool (image from [70]).

4.1.6 Discussion

Several databases, using a variety of computational methods to make inferences about functional relationships between genes and proteins, are available on the web [83–86]. HomoMINT is an inferred human protein network obtained by transferring the experimental interaction annotation from the proteome of seven model organisms to the corresponding ortholog human proteins. The orthology mapping is obtained by means of the InParanoid algorithm. Approximately one fifth of the interactions present in the MINT database could be mapped to human orthologs thus resulting in the assembly of an inferred network linking 4125 human proteins with 9749 edges. While a large proportion of these proteins are not functionally annotated one can use HomoMINT to transfer functional information from better characterized neighbors in the graph. Because of evolutionarily frequent molecular processes leading to gene family expansion or contraction, the transfer of interaction information between organisms, especially high eukaryotes, is complicated by the abundance of paralogs in orthology groups. The InParanoid algorithm is designed to distinguish paralogs arising before or after speciation events. We have chosen to transfer the interaction information only to the main human ortholog in each group. Thus our inferred network is essentially based on orthology mapping by the reciprocal best hit approach. However, the orthology groups assembled in our web available table contain paralogs, so permitting any alternative choice. Furthermore since the InParanoid algorithm provides a confidence score for each orthology assignment the likelihood of the inferred interactions can be evaluated from the confidence score of the model organism and human gene orthology assignment as proposed for instance by Lehner and Fraser [72]. To assess the predictive value of HomoMINT, we performed a number of tests aimed at assessing to

what degree of accuracy and coverage the orthology based inferred networks could be supported by previous knowledge. We first assembled a human experimental network from the protein interaction data stored in PPI databases and determined the percentage overlap between this network and HomoMINT or related networks. Next, we estimated the enrichment in the inferred networks of interacting proteins sharing Gene Ontology annotation. Finally we estimated the overlap between the inferred networks and the iHOP literature network. Our approach is based on the assumption that protein interactions between ortholog proteins are conserved in evolution. To what extent this is true cannot at present be estimated because of the incompleteness and inaccuracy of the available experimental datasets [64]. Even hypothesizing that the assumption is 100% correct, the accuracy and coverage of the inferred network is still limited by the quality of the original model organism interaction datasets and our ability to identify the true human orthologs of a model organism protein. Not surprisingly our benchmark tests show that accuracy increases if one uses more stringent criteria for orthology assignment (for instance by only allowing orthologs with similar modular architecture) or if one bases the inference on a more reliable interaction dataset (for instance relying on multiple evidence). In contrast with similar projects [72, 77, 87], HomoMINT is unique for its direct link to a curated PPI database. HomoMINT is a calculated section in the MINT relational database and its content is updated daily to take into account the newly curated entries in the MINT database. Furthermore the MINT viewer makes it possible to analyze and edit the HomoMINT network in the context of the experimentally verified interactions deposited in the MINT database. HomoMINT can be searched and analyzed at http://mint.bio.uniroma2.it/mint/search/search.php?dataset=homomint. The HomoMINT dataset is available either as a flat file or a PSI XML file, by trew bfollowing the link "download" at the same web address. Each of the provided files contains all interaction inferred from model organism's protein on main human orthologs.

Since it is not clear which percentage of PPI are conserved through evolution [64] HomoMINT should be considered as a hypothetical network that can be of use in predicting functions of yet uncharacterized proteins, in making experimentally testable hypotheses about new participants in well studied pathways and in prioritizing interactions to be tested in large scale PPI experiments. As such, the network should provide a rich source of functional hypotheses for researchers interested in the functions of one or many human proteins.

4.2 Results and discussion for the wiring of the yeast proteome by data integration approaches

4.2.1 A statistical framework for experimental technique independent high-throughput data assembling after a *data integration based* estimation of protein interaction

Due to the huge amount of protein interaction data, often incongruent, it would be desirable to develop a statistical framework able to include as many as possible different experimental technique results: this framework should be able to extract the information on the experimental technique specific random and systematic error contained within the high-throughput data sets itself. For example, maximum-likelihood estimation of experimental error rates permits examination of the self-consistency of data within a dataset [97]. Moreover the model should be able to produce a complete weighted graph of pairwise protein associations that can be converted into an unweighted graph (an high coverage approximation of the physiological yeast interactome), showing statistical properties consistent with that found in other studies of yeast protein interaction networks. It is possible to convert the graph into an unweighted graph by choosing a probability threshold, including only edges with weights above this threshold. More in detail, in the process of assembling TAP-MS data sets with two-hybrid assay data sets, it is important to figure out how to scale the evidences for direct interactions (two-hybrid assay results) into probabilities for two proteins that may be in the same complex but not interact directly. One way to realize this scaling is illustrated in the simple model below:

Given H_a, the hypothesis that two proteins, A and B, are in the same complex, let's imagine to have an evidence in two hybrid data (THD), that A and B interact directly, $A - B|THD$.

We can express the conditional probability as:

$Pr(H_a|twohybriddata(THD))$.

According to the hypothesis of a direct interaction:

$Pr(H_a|THD) = Pr(A-B|THD)$ where A-B indicates that A and B interact directly.

Let's consider now the indirect interaction scenario: If A and B are in the same complex but do not interact directly,

$P(H_a|THD) = P(A - C, C - B|THD)$

that, by assuming independence between experiments and that proteins can only be in one complex at a time, can be written as:

$P(A - C, C - B|THD) = P(A - C|THD)P(C - B|THD)$

It is possible to calculate these probabilities for all possible indirect interac-

tions (for example by building some kind of graph where the edges are weighted by the probability of any two proteins interacting). The knowledge of $H_a|THD$, allow us to combine information from two hybrid and TAP-MS data sets in a straightforward manner.

Next section faces the problem of finding an estimate of the probability of any two proteins to interact. This probability will be conceived as a system biology level probability as it will be calculated by modelling data coming from different experimental sources other than protein interaction.

4.2.2 Building a *data integration based* graph for protein interaction by modeling functional genomic and annotations evidences as predictors of protein interaction.

To calculate the probability of any two proteins interacting, a data integration approach has been followed, based on using a machine learning framework; a logistic regression model with some predictors derived from functional genomic data was trained on a huge and recent manually curated data set [91]. In the following sections the chosen genomic features and their modeling are described.

the mRNA co-expression feature COE

Expression data sources can be used for the prediction of protein interaction because proteins in the same complexe are often co-expressed [78, 79]. Modeling of expression level values in term of genomic feature, consisted in computing the Pearson correlation for all possible pairs in Rosetta compendium and cell cycle data sets, [80]. From the values related to all possible pairs, only those corresponding to the interacting pairs in the chosen gold standard were extracted.

the Co-abundance feature COABU

Protein abundance level data sources [94], can be modeled as a predictor of protein interaction because two interacting proteins should be present in stoichiometrically similar amounts. This predictor is simple the absolute value of the different concentration of the two proteins divided by the averaged concentration of proteins in the performed experiment [94].

The Gene Ontology based predictors: zscoreGOBP and zscoreGOCC

For each binary set $S(m)$ (i.e. the interacting protein pair) we computed the prevalence of all Gene Ontology (GO) terms among the annotated proteins in the binary set, and the probability that such prevalence would occur in a randomly chosen binary set of proteins. We always consider a protein/gene annotated

to a GO term if it is directly annotated to it or to any of its descendants in the GO graph. For a given GO term t let $K(t)$ be the total number of proteins/ORFs annotated to it in the proteome/genome, and $k(m,t)$ the number of proteins/ORFs annotated to it in the binary set $S(m)$. If J and $j(m)$ denote the number of proteins/ORFs in the proteome/genome and in $S(m)$ respectively, such probability is given by the right tail of the appropriate hypergeometric distribution (the so called Fisher's exact test):

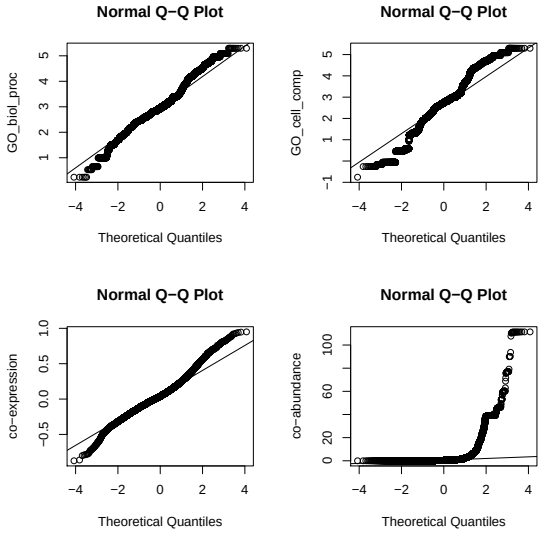
$$P(J, K(t), j(m), k(m,t)) = \sum_{h=k(m,t)}^{\min(j(m),K(t))} F(J, K(t), j(m), h) \tag{4.1}$$

where

$$F(K(t), k(m,t), J, j(m)) = \frac{\binom{K(t)}{k(m,t)}\binom{J-(K(t)}{j(m)-k(m,t)}}{\binom{J}{j(m)}} \tag{4.2}$$

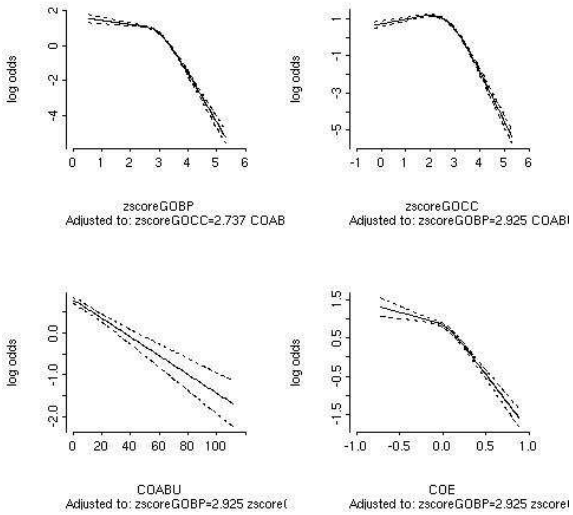
In this way a P-value can be associated to each pair made of an interacting pair and a Gene Ontology term. A low P-value indicates that the interaction is correlated to the functional characterization described by the GO term, and hence suggests the interaction a candidate for further experimental validation or for being included in a model as a predictor.

Checking the distributional assumptions of the values in the genomic features

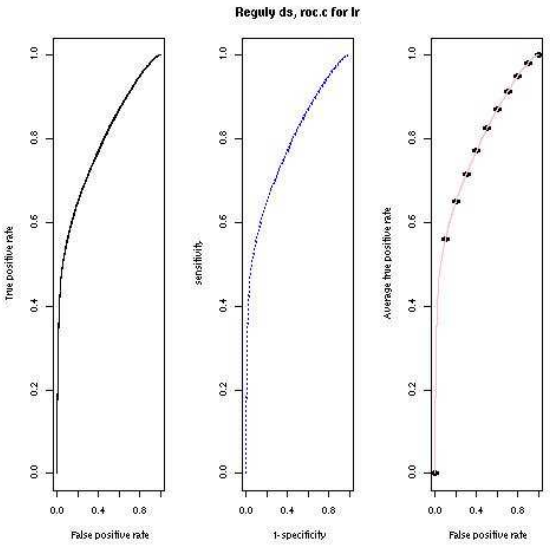


Comparison of the quantiles of the different genomic features against the quantiles from a Normal distribution.

Model evaluation and validation



On the y axis the odds ratio; on the x axis the predictor variables; the odds ratio can be considered a measure of association between the outcome (in this case the response variable Interacting class) and the predictor variable. The co-abundance predictor(COABU) for small values is more associated to the outcome yes.



4.2.3 Discussion

Several yeast predicted interactomes to make inferences about functional relationships between yeast proteins, are available on the web: they were built by

using a variety of genomic features and statistical methods [86, 93]. The model proposed here is also obtained by exploiting functional genomic data sets and functional annotation data sets. Even so, it doesn't want to represent another yeast predicted interactome but a piece of a bigger project: to build a software tool for the assembling of high-throughput data sets based on a hierarchical bayesian model with the peculiar bayesian *a priori* probability for two random proteins to interact, coming instead from a machine learning framework trained on recent completion of a comprehensive and systematic primary literature curation effort [91].

This effort , has made available a comprehensive database that currently houses a total of 22,250 protein interactions and 11,061 genetic interactions, corresponding to 11,334 and 8,165 nonredundant interactions in the so called Literature Curated Protein Interactome. This data set has been conceived as a look-up table for gene and protein interactions and as a basis for interrogating the properties of biological networks but in our knowledge, its potential in machine learning frameworks as training set has not yet been exploited.

4.3 Results and discussion for the wiring of the human proteome by evolutionary constraints

As a result of the selection pressure exerted on interacting proteins to mantain or modulate their capacity to interact during evolution, signals originated from coordinates changes in interacting protein sequences or in some their specific amino acid residues should exist. Generally, we refer to these signals as co-evolutionary signals . Similarity shown in evolutionary history of two proteins and patterns of covarying amino acids are the co-evolutionary signals examined in this part of the thesis with the final aim to verify if it is possible to use them as predictor of protein interaction.

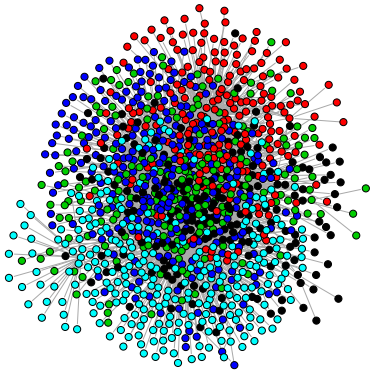
4.3.1 Correlated evolution analysis

One way to capture the evolutionary information is to look at the phylogenetic tree of the genomes [98–100]. Of particular interests is the mirror tree method [98]; it predicts protein protein interactions using the assumption that the interacting proteins show similarity in their respective molecular phylogenetic trees because of the co-evolution caused by the interaction. It is difficult, however, to directly evaluate the similarity between a pair of molecular phylogenetic trees. Instead, the mirror tree method compares a pair of distance matrices, from which the phylogenetic trees are reconstructed. Specifically, the extent of co-evolution between two proteins is measured as the element pairwise Pearson correlation between the two corresponding distance matrices.

In spite of its remarkable success, the mirror tree method suffers from a relatively high rate of false positives. Two proteins that do not interact may still have high Pearson correlation between their corresponding distance matrices. To address this issue, it was recently suggested by Sato et al. [101] that the information about the phylogenetic relationships of the analysed genomes should be excluded by a projection operation, and only the residual information in the distance matrices would be used for the calculation of the Pearson correlation coefficient. As a result, the false positive rate was indeed decreased, but so was the sensitivity. That is, as compared to the original mirror tree method, the Sato's method predicts more protein pairs as non-interacting, which are actually interacting pairs. Further improving have been recently proposed by Craig and Liao [102].

4.3.2 Definition of the subinteractome to evaluate in term of evolutionary constraints

We have chosen to analyse in term of evolutionary constraints that portion of the Human interactome called the *SH2ome*, a set of human interacting proteins in which at least one in every interacting protein pair is an *SH2 domain containig* protein. As a reference collection of the complete set of human proteins containing the peptide binding domain SH2 we have chosen the list provide by Pawson [60]. More details on the assembling of the *Human SH2ome* by using the available resources are provided in chapter 3.



4.3.3 Correlated evolution analysis of the *Human SH2ome*

A set of 3146 interacting protein pairs, resulting from the assembling of data from different protein interaction databases were the input for available bioinformatics pipelines (see chapter 3), aimed at obtaining for each protein in the network a corresponding multiple sequence alignment of the protein with its available orthologs in the Ensembl database [59]. The resulting set of 3146 pairs of multiple sequence alignments was the input for the mirror tree analysis, implemented as in [103]. The distribution of the correlation coefficient values generated by the mirror tree method is represented in form of histogram in the figure.

4.3.4 Definition and implementation of a null model for correlated evolution analysis

Let' s assume the *Human SH2ome* as a graph in which the edges connecting the SH2 domain containing proteins with the other nodes (protein binders) are weighted edges. The weight is represented by the correlation coefficient, output of the mirror tree approach that was previously applied recoursevely on every pair of multiple sequence alignments; we can immagine to collapse the reference protein of each multiple sequence alignment plus its orthologs in one node of the graph, representing the entire set of evolutionary related proteins.
In the following sections are reported the results for testing in order the following null hypothesis:

- The distribution of correlation coefficients in the complement graph is comparable to the distribution of correlation coefficients in the reference graph (the *Human SH2ome*).

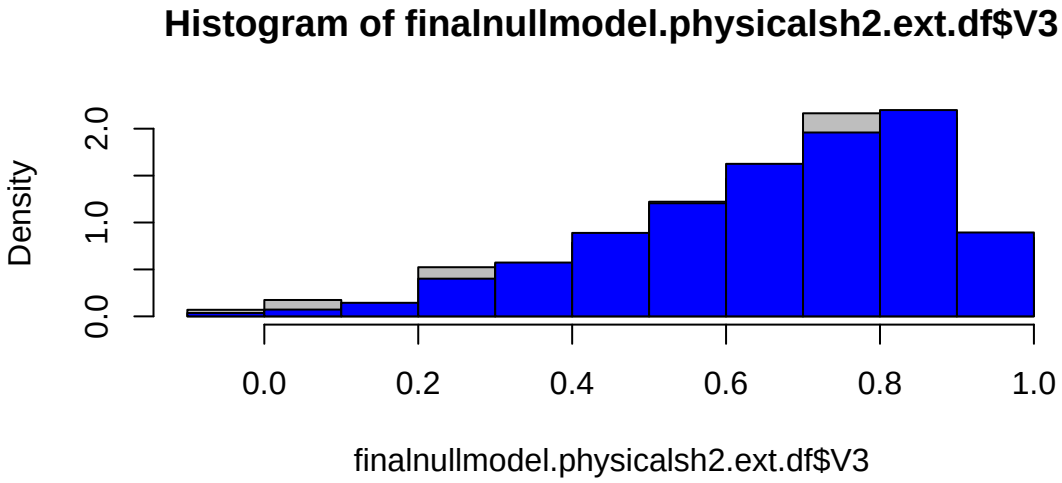
- following a binning step in which the reference graph (the *Human SH2ome*) was splitted in smaller subsets according to the node degree of the SH2 containing proteins, the distribution of correlation coefficients in the hub (high node degree) subsets of the reference network is comparable to the distribution of correlation coefficients in the graphs obtained by rewiring the same hubs with proteins never seen as binders in the reference set. We refer to this analysis as Hub - no Hub analysis.

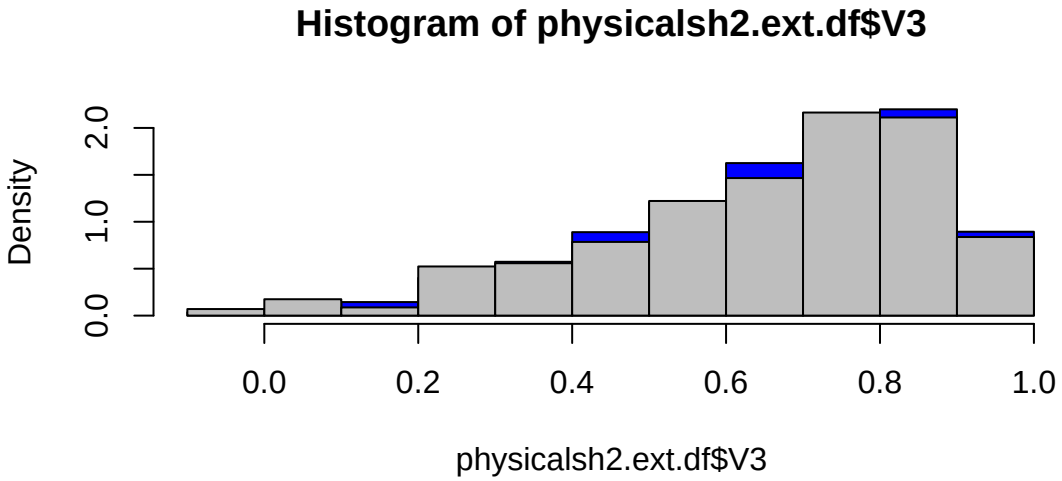
- following the same binning step previously described, the reference graph (the *Human SH2ome*) was trasformed in other sets of relationships according to the neighboorhood relationships between the SH2 containing proteins and the other nodes; it was evaluated if the distribution of correlation coefficients in the distance=2 sets in some hub proteins is comparable to the distribution of correlation coefficients in the distance=1 set (the reference graph).

- following the same binning step previously described, the reference graph (the *Human SH2ome*) was trasformed in other sets of relationships according to the neighboorhood relationships between the SH2 containing proteins and the other nodes; it was evaluated if the distribution of correlation coefficients in the distance=2 sets for some hub proteins is comparable to the distribution of correlation coefficients in a mixture of distances sets.

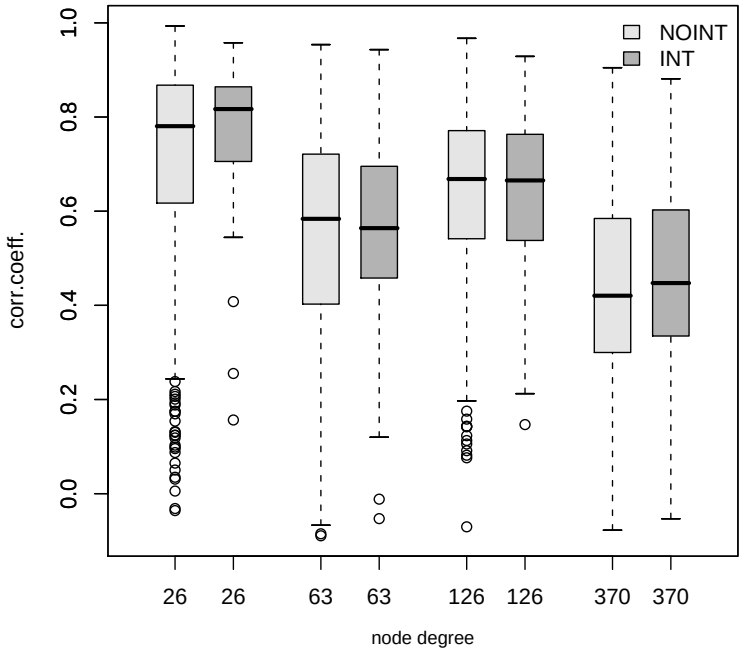
The complement graph of the *Human SH2ome*

For this analysis we used the same set of pairs of multiple sequence alignments rewired respect the reference set of interacting protein pairs. Here, a pair of multiple sequence alignments is considered as an interacting protein pair because the set of sequences constituting the multiple sequence alignment was collapsed in one protein, the human reference protein.
Only a sample of the huge amount of edges constituting the complement graph was considered for the analysis.



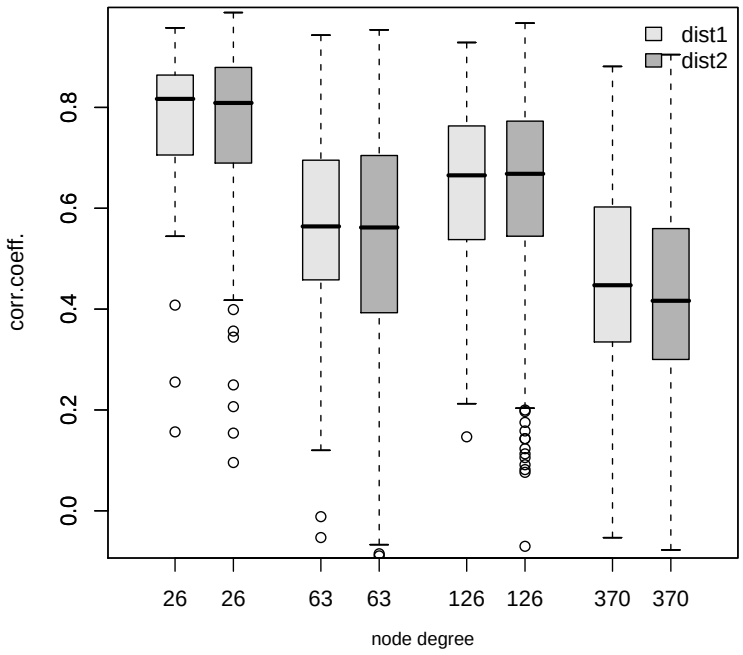


Hub - no Hub analysis

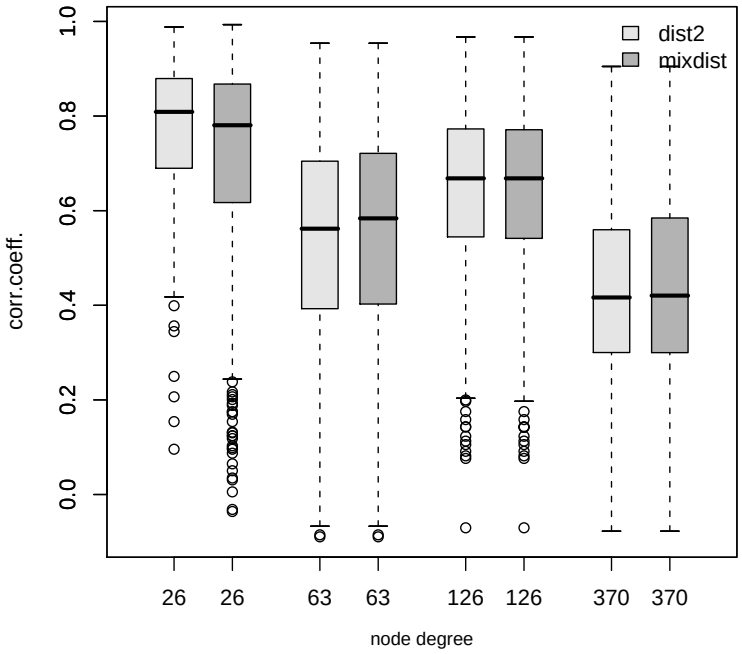


On the x axis, the node degree of the analysed proteins; on the y axis the correlation coefficient values, output of the mirror tree method. Correlation coefficients can be considered as the edge weights of the*Human SH2ome* network. The relation generating the weighted edges of the graph was the correlation coefficient between the distance matrices of set of evolutionary related proteins collapsed in the two reference physically interacting proteins. The same relation was applied in building an hypothetical graph linking the same hub proteins with a set of proteins d representing the list of proteins observed as non interacting with the hub in the original data set.

Distance or neighborhood analysis



On the x axis, the node degree of the analysed proteins; on the y axis the correlation coefficient values, output of the mirror tree method. Correlation coefficients can be considered as the edge weights of the *Human SH2ome* network. The relation generating the weighted edges of the graph was the correlation coefficient between the distance matrices of set of evolutionary related proteins collapsed in the two reference proteins at distance 2 in the graph. The same relation was applied in building a graph linking the same hub proteins with a set of proteins d representing a list of proteins observed at distance 1 from the analysed hubs in the original data set. This means that the distance 2 graph was compared with the reference graph.



On the x axis, the node degree of the analysed proteins; on the y axis the correlation coefficient values, output of the mirror tree method. Correlation coefficients can be considered as the edge weights of the*Human SH2ome* network. The relation generating the weighted edges of the graph was the correlation coefficient between the distance matrices of set of evolutionary related proteins collapsed in the two reference proteins at distance 2 in the graph. The same relation was applied in building an hypothetical graph linking the same hub proteins with a set of proteins d representing a list of proteins observed at various distances from the analysed hubs in the original data set.

4.3.5 Protein evolutionary covariation analysis

The other co-evolutionary signals examined in this part of the thesis are the signals originated from coordinates changes of amino acid residues. According to the definition of Hakes et al. [104] these changes are a direct result of maintaining optimal structural and functional integrity of the interacting proteins and can be considered a form of coadaptation. More in detail we refer to the analysis aimed at the detection of these kind of signals as Protein Evolution Covariation Analysis (PECA) [105]. PECA is also the suite of perl programs and modules implementing this kind of analysis, aimed at measuring co-evolution events between amino acid sites within (intramolecular co-evolution) or between (inter-protein co-evolution) proteins. Essentially the program performs the detection of amino acid sites undergoing co-evolution from multiple sequence alignments (MSAs). In addition, PECA detects groups of co-evolution using several statistical filters, including parsimony-informative pairs of coevolving sites, sites with correlated variation in their hydrophobicity and/or in their molecular weight characteristics. These filters are used as a priori tools for the identification of biologically meaningful coevolving pairs of sites. This filtering approach ensures an increase in the accuracy from 17% [107] to 85% [106].
To test the hypothesis of coevolution between proteins, PECA uses the non-parametric method based on the Information Theory and on the Mutual Information Content (MIC)(see Chapter 2). According to this method, developed by Korber and colleagues [109], mutual information is represented by the entropies that involve the joint probability distribution, $p(s_i, s_j^{'})$, of occurrence of symbol i at position s and j at position s' of the multiple sequence alignment(see Chapter 2). The MIC values generated range between 0, indicating independent evolution, and a positive value whose magnitude depends on the amount of covariation. Variable positions included in the alignment and considered in the coevolutionary analyses were those parsimony-informative (i.e. they contain at least two types of amino acids and at least two of them occur with a minimum frequency of two). The significance of the MIC values was assessed by randomization of pairs of sites in the alignment, calculation of their MIC values and comparison of the real values with the distribution of one million randomly sampled values. PECA has been applied to the same set of multiple sequence alignments interacting proteins examined in the paragraph 4.3.2 with the initial aim to verify if there is some correlation between the *node degree* of SH2 containing proteins in the network under analysis, the correlation coefficient calculated in section 4.3.2 and the number of co-evolving residues detected by using the inter-protein co-evolution analysis option implemented in PECA. The results of this analysis are summarized in the table 4.5.

Does covariation can be a predictor for protein interaction?

			Binder name	corr.coeff (mirrortree)
	Nr. of coev. res pairs	Nr. of groups		
SOCS-3, degree 26	2	2	ENSG00000100181	0.156559143928572
	4	1	ENSG00000145715	0.544341963045092
	13	5	ENSG00000162434	0.957774084014608
CRK, degree 63	0	0	ENSG00000107643	-0.0529964795716593
	21	3	ENSG00000146648	0.510916555565708
	2	1	ENSG00000141736	0.889452028412915
	0	0	ENSG00000135407	0.94319862489751
			Binder name	**corr.coeff (mirrortree)**
	Nr. of coev. res pairs	Nr. of groups		
GRB1, degree 126	7	2	ENSG00000167193	0.146763593678151
	18		ENSG00000165025	0.510624502399393
	13	1	ENSG00000141736	0.751034572987499
	9	2	ENSG00000137462	0.928739441131793
ASH, degree 370	0	0	ENSG00000121774	-0.0531852627891199
	57	7	ENSG00000146648	0.522278435163278
	6	2	ENSG00000110700	0.881229058180765

Table 4.5: Table showing the relationships between number of coevolved residue pairs and correlated evolution (measured by the correlation coefficient according to the mirrotree approach [98]. The results are binned according to the node degree values for the *Human SH2ome* proteins analyzed.

In a very recent work, Yeang et al. found that a very small fraction of physical interactions (2.6%) from Pfam [110] scored significantly for covariation. The conclusion of this large scale screening on all the known protein families are that physically interacting amino acid residues are not coevolved and that covariation seems not necessary for physical interactions, being these dominated by conserved or with unilateral changes sequences. Nevertheless, coevolution can manifests spatial and functional constraints other than direct interactions that can be captured in the interacting proteins looking not at the interaction surfaces but at the regions surrounding the patches involved in the physical binding. A predictor for protein interaction based on coevolving amino acid sites can be built by performing as first step an intramolecular co-evolution analysis aimed at identifying coevolutionary patterns; a second step intermolecular analysis can be performed restricted on the previously identified regions. To apply this framework to biological relevant networks, we have to deal with the problem of making the coevolutionary results comparable between the different interacting pairs.

The strength of the coevolutionary pattern can be calculated by classifying significant MIC values into categories (i.e. 0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, $MIC > 0.50$); 0.1 can include all those pairs of amino acid sites with MIC values $0 < MIC0.1$; 0.15 would include $0.1 < MIC0.15$, and so

on. Actually, this categorization of MIC values allows the direct comparison of the coevolutionary results between different pairs of proteins regardless the set of MIC values obtained in each analysis: to quantify the contribution of each category to the overall MIC value, we need first to count the number of pairs of sites showing MIC values within a certain category. Then, we need to calculate the percentage of pairs of sites included in that category by dividing the number of sites in the category by the total number of pairs of sites detected as coevolving significantly. This way, the contribution of each MIC category between pairs of proteins is comparable. Starting from these considerations it is straightforward and exciting to end up with a new predictor of protein interaction purely information content and genomic information based.

Chapter 5

Overall discussion and perspectives

Protein interaction networks derived from experiments. The fast development of experimental techniques for protein interactions has enabled the construction and systematic analysis of interaction networks [89,90]. As described in sections 3.1 and 4.1, interaction maps obtained for one species can be used to predict interaction networks in other species, to identify functions of unknown proteins, and to get insight into the evolution of protein interaction patterns. The comparisons of interaction maps are based on the observation that many interactions are conserved among species (interologs) [81]. Sequence-based searches for interologs were able to identify from 16% to 31% of true interologs (tested using Y2H system) even between remotely related species such as yeast and worm [82]. Moreover, analysis of conservation in the networks produced by gene co-expression data revealed that interologs correspond to the functionally related genes responsible for core biological processes [111].
Nevertheless, signaling events regulated by tyrosine phosphorylation and dephosphorylation typical of human cells do not take place in yeast or in the other interactomes used to infer the predicted human interactome; due to significant differences in genome size and protein architecture in human and yeast, most human protein interactions, especially those occurring in signal trasduction and cell-cell communication may not be directly inferred from evolutionary distant interactomes. Indeed a comparative analysis of more than 70.000 protein interactions identified to date for yeast, worm, fly, and human, revealed that only 16 interactions are common to all the four species suggesting that network of protein interaction are species specific or differ significantly from on specie to the other [112]. Moreover, systematic identification of direct protein-protein interactions is often hampered by difficulties in expressing and purifying the corresponding full lenght proteins in higher organisms. For all these reasons, the

development of low cost and efficient methods for studying protein interaction represents a step very important in consolidating our knowledge, in prioritizating the experimental efforts and in giving value to the huge amount of data still incongruent. It will be important in the future to try to develop prediction methods able to exploit very different kind of data with statistical frameworks able to extract the error inherent in a specific data set and to collocate the different experimental sets in a common space; it will be very important also to take in account the biology of a system. In a recent analyses of high-throughput protein interaction data coupled with investigations of evolutionary properties, Mintseris et all. [113] pointed to the importance of distinguishing between transient and obligate interactions. The time seems to be a variable very important to take in account: residues in the interfaces of obligate complexes tend to evolve at a relatively slower rate, allowing them to coevolve with their interacting partners. The mutations have possibility to be fixed in the system; the plasticity inherent in transient interactions leads to an increased rate of substitution for the interface residues and leaves little or no evidence of correlated mutations across the interface.

Conclusion

In this work, we have presented some computational frameworks, developed for wiring human and yeast proteomes on a genome-wide scale. The proposed methodologies are based on several ingredients, namely statistical modeling approaches to integrate different kind of experimental data, evolutionary constraints, different kind of functional annotations. Moreover, all the proposed workflows are able to give biological meaningful results. Their application to the yeast and to the human case are reported, with discussion of and future perspectives.

Appendix A

Biological glossary

In this chapter, we report some useful biological definitions.

coding sequence a DNA sequence that encode for a gene product such as a protein.

coding strand the strand of the DNA duplex that is really translated into a protein through the genetic code.

codon a trinucleotide (triplet, or 3-word) that specifies for a particular amminoacid through the genetic code.

differential expression the expression of one or more genes to different extents, depending upon growth conditions, treatments applied, or the state of the cell cycle.

euchromatin the portion of a chromosome that is less condensed and more transcriptionally active.

eukaryote organisms characterized by a true membrane-bounded nucleus containing chromosomes complexed with histones, a cytoskeleton, and mebrane-bound organelle such as mithocondria. Humans and yeasts are eukaryotes.

exon a contigous segment of DNA that is represented in a processed mature RNA molecule after splicing has removed intronic sequences. Exon sequences may be translated or untranslated.

gene a genomic locus (or DNA segment) specifying or contributing to an heritable trait associated with an organism. A gene usually encode for a RNA species and (after translation) to polypeptide chains, the proteins.

genome the entire genetic complement of an organism.

homologs related genes or loci whose similarity is a consequence of descent from a shared common ancestor. Homologs in different species are *orthologs*, and homologs within a species are *paralogs*.

intron a segment of non-coding DNA separating exons within genes. Introns are removed by splicing from precursor RNA molecule to form mature mRNA.

locus a position on a genetic map or genome defined by a certain gene or DNA sequence appearing at that position.

mutation a heritable change in DNA relative to a defined "wild-type" reference sequence.

non coding sequence DNA sequence that does not appear in the final gene product. This include intergenic sequences, introns, untraslated regions.

open reading frame *ORF* a succession o triplet not interrupted by STOP codons.

ortholog homologs appearing in different species.

paralog homolog arising from a gene duplication within a single lineage or species instead of arising by descent in diverging lineages.

prokaryote an organism that does not contain a true nucleus, membrane-bound organelle, or complex cytoskeleton.

promoter a DNA sequence element required for initiation of transcription of a gene, including sites where transcription factors bind to control the time and cell type in which transcription occurs.

proteome the complete set of protein encoded by a certain organism.

repeated sequence a DNA sequence that appears more than once in a genome.

splicing processing of primary RNA transcript to remove introns and to produce mature mRNA molecules containing a continuous coding sequence composed by joined exons.

spottet microarray a collection of DNA probes, used for measuring gene expression levels.

transcription factors proteins involved in the starting of transcription for a certain gene trough binding to specific loci in the promoters regions (transcription factor binding sites, TFBSs) of the regulated gene.

transcriptome the complete collection of mature transcripts in a particular cell type under a specified set of physiological and environmental conditions.

Appendix B

Bioinformatic glossary

In this chapter, we report some useful bioinformatics definitions:

alignment the procedure by which two (or more) nucleic (or protein) sequences are arranged to establish a relationship between them.

binomial distribution the binomial is the probability distribution describing the number of successes and failures in a fixed number of independent trials when only two outcomes are possible, often called "success" and "failure". The number of heads in some fixed number of tosses of a coin is an example of a binomial random variable. If we denote with Y the total number of success in n trials, so the probability distribution of Y is given by the formula

$$P_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y = 0, 1, 2, ..., n.$$

where:

$$\binom{n}{y} = \frac{n!}{(n-y)!y!}$$

BLAST *Basic Local Alignment and Search Tool* a program for rapidly searching protein or DNA sequences in a database and to detect statistically significant local alignments throught heuristic procedures.

Bonferroni correction the bonferroni is a multiple testing correction method. If N multiple independent hypotheses are tested for significance, the correct $cutoff$ should be written as $\simeq \frac{cutoff}{N}$.

clustering the procedure of grouping together objects based upon some kind of similarities or distance measure between them.

consensus sequence a short DNA or protein sequence having, at each position, the most probable letter at that position.

conserved sequence a DNA or protein region found nearly identical in two or more genomes, after alignment.

deletion an alteration in DNA sequences resulting from removing one or more contiguous bases from the sequence string.

dissimilarity a measure of the degree of difference between objects with respect to a certain distance measure.

false discovery rate or *FDR*, in classification, the fraction of those features identified as positive that are in fact false positives.

FASTA a rapid local alignment method based upon locations of k-words in a alignment matrix. This allow a more detailed examination of regions where hits are frequent.

gene expression matrix for n genes whose expression is measured for m conditions, the $n \times m$ matrix of expression levels (typically ratios of treatment to control conditions) is the gene expression matrix.

global alignment alignment between two sequences such that all letters of both sequences are aligned opposite letters or indels.

Gold Standard it is a data set consisting of a number of known interacting and non-interacting protein pairs, which are used to train classifiers and estimate their predictive ability.

hit a database entry matching a query sequence after a database search.

hypergeometric distribution suppose that an urn contains N objects, of which n are red and $N - n$ are white. Of these, m objects are taken out of the urn at random, in particular without reference to the color and without replacement. The number of red objects taken out is a random variable Y, with probability distribution given by the formula

$$P_Y(y) = \frac{\binom{n}{y}\binom{N-n}{m-y}}{\binom{N}{m}} \quad y = A, A+1, ..., B.$$

where $A = max(0, n + m - N), B = min(n, m)$.

indel an insertion or deletion of letters applied to either of two sequences string being aligned.

A = Adenine	**R = A or G (purine)**	**M = A or C**
C = Cytosine	**Y = T or C (pyrimidine)**	**B = T,G or C**
G = Guanine	**S = G or C**	**V = A,G or C**
T = Thymine	**W = A or T**	**H = A,T or C**
U = Uracil	**K = G or T**	**F = A,T or G**
	N = any base	

IUPAC-IUB symbols symbols for DNA combination bases:

information of a sequence: a measure of it nonrandomness. Can be measure *i.e.* using relative entropy or Shannon's entropy.

insertion the addition of one or more nucleotides into a nucleic acid sequence.

interactomes descriptions of the functional connections between a variety of different biological entities, metabolites linked by enzymatic reactions (metabolic networks), transcriptions factors linked to DNA binding sites (genetic networks); in this thesis, proteins functionally or physically linked to other proteins in the protein networks. In literature, the results of systematic experiments performed under physiological conditions or the compilation and the assembling of partial data inferred from experiments performed under a variety of different conditions.

local alignment alignment of substrings taken from each of two different sequence strings.

machine learning see also supervised-unsupervised.

motif a short local sequence pattern found among a set of proteins or DNA sequences.

multiple alignment alignment of more than two sequences strings.

multiple hypothesis testing the simultaneous testing of two or more alternative hypotheses.

pairwise alignment alignment of two sequence strings.

PAM or *Point Accepted Mutations* a set of matrices for scoring amino acid or DNA substitutions in alignments.

phylogenetic footprinting a DNA sequence pattern recognized to appear similar in aligned regions of related genomes.

position specific scoring matrix or *PSSM* a matrix whose rows correspond to letters that occur at positions in a DNA signal and whose columns corresponds to the positions. Matrix element are the log-odds scores for each letter at each position, computed relative to an appropriate null model. A PSSM is a particular type of a PWM.

positional weight matrix or PWM a matrix whose rows correspond to letters that occur at positions in a DNA signal and whose columns corresponds to the positions. Elements of this matrix are related to the probability of occurrence of each letter at each position.

substitution matrix a matrix specifying scores to be applied for matching DNA sequences in an alignment. An example is the PAM matrix.

supervised-unsupervised want to learn an unknown function $x = y$, where x is an input example and y is the desired output

Appendix C

Publications

Publications directly related to the thesis work:

- [70] Persico M., Ceol A. , Gavrila C. , Hoffmann R. , Florio A. and Cesareni G.
  HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms .
  BMC Bioinformatics December 6(Suppl 4):S21

- [64] Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, Schneider MV
  Comparative interactomics.
  FEBS Lett 2005, 579:1828-33.

- Persico M., Hoffmann R., Capobianco E., Marras E.
  Evolutive constraints for "wiring" and characterizing human subproteomes
  Poster at BITS2007 Meeting, 26-28 April, Naple, Italy

Publications not directly related to the thesis work:

- Francesca Granucci, Caterina Vizzardelli, Norman Pavelka, Sonia Feau, Maria Persico, Ettore Virzi, Maria Rescigno, Giorgio Moro and Paola Ricciardi-Castagnoli
  Inducible IL-2 production by dendritic cells revealed by global gene expression analysis.
  nature immunology, volume 2 no 9, september 2001.

Bibliography

[1] Fields S, Song O.
A novel genetic system to detect protein-protein interactions.
Nature. 1989 Jul 20;340(6230):245-6.

[2] Causier B. *Studying the interactome with the yeast two-hybrid system and mass spectrometry.*
Mass Spectrom Rev. 2004 Sep-Oct;23(5):350-67. Review

[3] Walhout AJ, Vidal M.
High-throughput yeast two-hybrid assays for large-scale protein interaction mapping.
Methods. 2001 Jul;24(3):297-306.

[4] Lee JW, Lee SK.
Mammalian two-hybrid assay for detecting protein-protein interactions in vivo.
Methods Mol Biol. 2004;261:327-36. Review.

[5] Bartel PL, Roecklein JA, SenGupta D, Fields S.
A protein linkage map of Escherichia coli bacteriophage T7.
Nat Genet. 1996 Jan;12(1):72-7.

[6] Deeds EJ, Ashenberg O, Shakhnovich EI.
A simple physical model for scaling in protein-protein interaction networks.
Proc Natl Acad Sci U S A. 2006 Jan 10;103(2):311-6.

[7] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, 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 AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Srensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M.
Systematic identification of protein complexes in Saccharomyces cerevisiae

by mass spectrometry.
Nature. 2002 Jan 10;415(6868):180-3.

[8] Gavin AC, Bsche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hfert 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 MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G.
Functional organization of the yeast proteome by systematic analysis of protein complexes.
Nature. 2002 Jan 10;415(6868):141-7.

[9] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrn-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF.
Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.
Nature. 2006 Mar 30;440(7084):637-43.

[10] Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, et al.
A generic protein purification method for protein complex characterization and proteome exploration.
Nat Biotechnol 1999 17: 10301032.

[11] Di Tullio A, Reale S, De Angelis F.
Molecular recognition by mass spectrometry.
J Mass Spectrom. 2005 Jul;40(7):845-65. Review.

[12] Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB.
Electrospray interface for liquid chromatographs and mass spectrometers.
Anal Chem. 1985 Mar;57(3):675-9.

[13] Pieles U, Zrcher W, Schr M, Moser HE.
Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides.
Nucleic Acids Res. 1993 Jul 11;21(14):3191-6.

[14] Venable JD, Xu T, Cociorva D, Yates JR 3rd.
Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra.
Anal Chem. 2006 Mar 15;78(6):1921-9.

[15] *Open mass spectrometry search algorithm.*
J Proteome Res. 2004 Sep-Oct;3(5):958-64.

[16] Piehler J.
New methodologies for measuring protein interactions in vivo and in vitro.
Curr Opin Struct Biol. 2005 Feb;15(1):4-14.Click here to read

[17] Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al.
Global analysis of protein activities using proteome chips.
Science 2001 293: 21012105.

[18] Margittai M, Widengren J, Schweinberger E, Schrder GF, Felekyan S, Haustein E, Knig M, Fasshauer D, Grubmller H, Jahn R, Seidel CA.
Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1.
Proc Natl Acad Sci U S A. 2003 Dec 23;100(26):15516-21.

[19] Jones RB, Gordus A, Krall JA, MacBeath G
A quantitative protein interaction network for the ErbB receptors using protein microarrays.
Nature 2006 439: 168174.

[20] Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, et al.
Gene function prediction from congruent synthetic lethal interactions in yeast.
Mol Syst Biol (2005) 1: 0026.

[21] Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, et al.
The Protein Data Bank and the challenge of structural genomics.
Nat Struct Biol 2000 7(Supplement): 957959.

[22] Smith GP
Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface.
Science (1985) 228: 13151317.

[23] Yan Y, Marriott G
Analysis of protein interactions using fluorescence technologies.
Curr Opin Chem Biol (2003) 7: 635640.

[24] Cooper MA
Label-free screening of bio-molecular interactions.
Anal Bioanal Chem (2003) 377: 834842

[25] Yang Y, Wang H, Erie DA
Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy.
Methods (2003) 29: 175187.

[26] Baumeister W, Grimm R, Walz J
Electron tomography of molecules and cells.
Trends Cell Biol (1999) 9: 8185.

[27] *http://dip.doe-mbi.ucla.edu/*

[28] *http://www.ebi.ac.uk/intact/*

[29] *ttp://mips.gsf.de/genre/proj/mpact*

[30] *http://www.thebiogrid.org/*

[31] *http://www.hprd.org/*

[32] *http://mips.gsf.de/genre/proj/dima2/*

[33] *http://mint.bio.uniroma2.it/mint/Welcome.do*

[34] *http://www.sanger.ac.uk/Software/Pfam/iPfam/*

[35] *http://bond.unleashedinformatics.com/Action?*

[36] *http://string.embl.de/*

[37] Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al.
The HUPO PSIs molecular interaction formatA community standard for the representation of protein interaction data.
Nat Biotechnol 22: 177183.2004

[38] Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stmpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H.
The minimum information required for reporting a molecular interaction experiment (MIMIx).
Nat Biotechnol. 2007 Aug;25(8):894-8.Click here to read

[39] *(http://imex.sourceforge.net).*

[40] Arnaud Ceol, Andrew Chatr-aryamontri, Elena Santonico, Roberto Sacco, Luisa Castagnoli, Gianni Cesareni
DOMINO: a database of domain-peptide interactions.
Nucleic Acids Research, 2007, Vol. 35, Database issue D557-D560

[41] Finn RD, Marshall M, Bateman A.
iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.
Bioinformatics (2005) 21:410412

[42] Pagel P, Oesterheld M, Stmpflen V, Frishman D.
The DIMA web resource exploring the protein domain network.
Bioinformatics (2006) 22:997998.

[43] Green DM, Swets JA.
Signal Detection Theory and Psychophysics.
Wiley, 1974.

[44] Bradley AP.
The use of the area under the ROC curve in the evaluation of machine learning algorithms.
Pattern Recognition 1997;30(7):114559.

[45] Metropolis N, Rosenbluth AW, Teller AH, et al.
Equations of state calculations by fast computing machines.
JChemPhys 1953;21:108791.

[46] Kirkpatrick S, Gelatt CD, Jr, Vecchi MP.
Optimization by simulated annealing.
Science 1983;220:67180.

[47] Glover F.
Future paths for integer programming and links to artificial intelligence.
Computers and Operations Research 1986;5:53349.

[48] Goldberg D.
Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[49] Koza JR. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* Cambridge, MA:The MIT Press, 1992.

[50] Larranaga P, Lozano JA (eds).
Estimation of Distribution Algorithms. A NewTool for Evolutionary Computation.
Boston Dordrecht London: Kluwer Academic Publishers, 2002.

[51] Thomas M. Cover and Joy A. Thomas.
Elements of Information Theory
. John Wiley & Sons, Inc., N. Y., 1991.

[52] D. K. C. MacDonald.
Information Theory and Its Applications to Taxonomy
. J. Applied Phys., 23:529–531, 1952.

[53] M. Tribus.
Thermostatics and Thermodynamics
. D. van Nostrand Company, Inc., Princeton, N. J., 1961.

[54] Benjamin A. Shoemaker, Anna R. Panchenko
Deciphering ProteinProtein Interactions. Part I. Experimental Techniques and Databases .
PLoS Comput Biol 3(3):e42. doi:10.1371/journal.pcbi.0030042

[55] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al.
A comprehensive analysis of proteinprotein interactions in Saccharomyces cerevisiae.
(2000) Nature 403: 623627.

[56] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al.
A comprehensive two-hybrid analysis to explore the yeast protein interactome.
(2001) Proc Natl Acad Sci U S A 98: 45694574.

[57] Barabasi AL, Albert R.
Emergence of scaling in random networks.
Science. 1999 Oct 15;286(5439):509-12.

[58] Kevin R Brown and Igor Jurisica
Unequal evolutionary conservation of human protein interactions in interologous networks.
Genome Biology 2007, 8:R95

[59] Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T et al.
Ensembl 2006.
Nucleic Acids Res. 34 D556-561 (2006).

[60] Bernard A. Liu, Karl Jablonowski, Monica Raina, Michael Arce, Tony Pawson, and Piers D. Nash
The Human and Mouse Complement Resource of SH2 Domain ProteinsEstablishing the Boundaries of Phosphotyrosine Signaling Molecular Cell 22, 851868, June 23, 2006

[61] BTM Korber, RM Farber, DH Wolpert, and AS Lapedes
Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type 1 Envelope Protein: An Information Theoretic Analysis
PNAS, Aug 1993; 90: 7176 - 7180.

[62] William R. Atchley, Kurt R. Wollenberg, Walter M. Fitch, Werner Terhalle and Andreas W. Dress
Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis
Molecular Biology and Evolution 17:164-178 (2000)

[63] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ
Basic local alignment search tool. J Mol Biol 1990, 215:403-10

[64] Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, Schneider MV
Comparative interactomics.
FEBS Lett 2005, 579:1828-33.

[65] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.
Bioconductor: open software development for computational biology and bioinformatics.
Genome Biol 2004, 5:R80

[66] Gusfield D.
Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.
Cambridge: Cambridge University Press; 1997.

[67] Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P.
SMART 4.0: towards genomic data integration.
Nucleic Acids Res 2004, 32:D142-4.

[68] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al.
The Pfam protein families database.
Nucleic Acids Res 2004, 32:D138-41

[69] Bairoch A.
The ENZYME database in 2000
Nucleic Acids Res 2000, 28:304-5

[70] Persico M., Ceol A. , Gavrila C. , Hoffmann R. , Florio A. and Cesareni G.
HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.
BMC Bioinformatics December 1;6(Suppl 4):S21

[71] Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M.
Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.
Genome Res. 2004 Jun;14(6):1107-18.

[72] Lehner B, Fraser AG.
A first-draft human protein-interaction map.
Genome Biol 2004, 5:R63.

[73] Remm M, Storm CE, Sonnhammer EL.
Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.
J Mol Biol 2001, 314:1041-52.

[74] Hegyi H, Gerstein M.
Annotation transfer for genomics: measuring functional divergence in multi-domain proteins.
Genome Res 2001, 11:1632-40.

[75] Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA.
Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 2004, 14:208-16.

[76] Hoffmann R, Valencia A
A gene network for navigating the literature. Nat Genet 2004, 36:664.

[77] Kevin R Brown and Igor Jurisica
Online Predicted Human Interaction Database.
Bioinformatics 2005, 21:2076-2082.

[78] Ge, H., Liu, Z., Church, G.M., and Vidal, M.
Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.
Nat. Genet. 2001, 29: 482-486.

[79] Gerstein, M., Lan, N., and Jansen, R.
Proteomics. Integrating interactomes.
Science 2002, 295: 284-287.

[80] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al.
A genome-wide transcriptional analysis of the mitotic cell cycle.
Mol. Cell 1998, 2: 65-73.

[81] et al.
Protein interaction mapping in C.elegans using proteins involved in vulval development.
Science 7 January 2000, vol 287. no. 5450 : 116-122.

[82] Matthews L. R. et al.
Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or interologs.
Genome Res. 2001 11:2120-2126.

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

[84] Huang TW, Tien AC, Huang WS, Lee YC, Peng CL, Tseng HH, Kao CY, Huang CY
POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome.
Bioinformatics 2004, 20:3273-6.

[85] Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C
Predictome: a database of putative functional links between proteins.
Nucleic Acids Res 2002, 30:306-9.

[86] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B
STRING: a database of predicted functional associations between proteins.
Nucleic Acids Res 2003, 31:258-61.

[87] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P
STRING: known and predicted protein-protein associations, integrated and transferred across organisms.
Nucleic Acids Res 2005, 33:D433-7.

[88] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G.
MINT: the Molecular INTeraction database.
Nucleic Acids Res. 2007 Jan;35(Database issue):D572-4. Epub 2006 Nov 29.

[89] Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al.
A protein interaction map of Drosophila melanogaster.
Science 2003, 302:1727-36.

[90] Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, et al.
A map of the interactome network of the metazoan C. elegans.
Science 2004, 303:540-3.

[91] Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M.
Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.
J Biol. 2006 Jun 8;5(4):11.

[92] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M.
BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535-9.

[93] Lu L.J., Xia Y., Paccanaro A., Yu H., Gerstein M.
Assessing the limits of the genomic data integration for predicting protein networks.
Genome Res. 2005 15:945-953.

[94] Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, OShea EK, Weissman JS
Global analysis of protein expression in yeast
Nature 2003, 425:737-741.

[95] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., et al.
survey Nature 2003, 425, 686-691.

[96] Ben-Hur A, Noble WS.
Choosing negative examples for the prediction of protein-protein interactions.
BMC Bioinformatics. 2006 Mar 20;7 Suppl 1:S2.

[97] Gilchrist MA, Salter LA, Wagner A.
A statistical framework for combining and interpreting proteomic datasets. Bioinformatics. 2004 Mar 22;20(5):689-700. Epub 2004 Jan 22.

[98] Florencio Pazos and Alfonso Valencia. *Similarity of phylogenetic trees as indicator of protein interaction* Protein Eng. 2001 Sep;14(9):609-14.

[99] Jothi R, Kann MG, Przytycka TM. *Predicting protein-protein interaction by searching evolutionary tree automorphism space.* Bioinformatics. 2005 Jun;21 Suppl 1:i241-50.

[100] Craig RA, Liao L. *Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices.* BMC Bioinformatics. 2007 Jan 9;8:6.

[101] Sato T, Yamanishi Y, Kanehisa M, Toh H.
The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics. 2005 Sep 1;21(17):3482-9. Epub 2005 Jun 30

[102] ROGER A. CRAIG AND LI LIAO
Improving Protein-Protein Interaction Prediction based on Phylogenetic Information using Least-Squares SVM
Ann N Y Acad Sci. 2007 Oct 9;

[103] Jothi R, Cherukuri PF, Tasneem A, Przytycka TM.
Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions J Mol Biol. 2006 Sep 29;362(4):861-75.

[104] Luke Hakes, Simon C. Lovell, Stephen G. Oliver, and David L. Robertson
Specificity in protein interactions and its relationship with sequence diversity and coevolution.
PNAS May 8, 2007 vol. 104 no.19 7999 8004

[105] David McNally and Mario A Fares
In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae
BMC Evolutionary Biology 2007, 7:81

[106] Codoer, F.M., ODea, S., Fares, M.A.
Using biological filters to improve the sensitivity of non-parametric methods to detect molecular coevolution
Mol Biol. Evol., in Press

[107] Fares, M.A. and Travers, S.A.
A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses.
Genetics (2006) 173, 9-23.

[108] Martin LC, Gloor GB, Dunn SD, Wahl LM.
Using information theory to search for co-evolving residues in proteins.
Bioinformatics. 2005 Nov 15;21(22):4116-24.

[109] Korber BT, Farber RM, Wolpert DH, Lapedes AS
mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A 1993, 90(15):7176-7180.

[110] Yeang C.H. and Haussler D.
Detecting coevolution in and among protein domains.
Plos Comp. Biol. Nov.2007 vol.3 issue 11 e211.

[111] Stuart JM, Segal E, Koller D, Kim SK.
A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003 Oct 10;302(5643):249-55.

[112] Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A.
Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.
Nat Genet. 2006 Mar;38(3):285-93.

[113] Mintseris J, Weng Z.
Structure, function, and evolution of transient and obligate protein-protein interactions.
Proc Natl Acad Sci U S A. 2005 Aug 2;102(31):10930-5.