University of Rome “Tor Vergata”

COMPARATIVE GENOMIC APPROACH FOR PROTEIN-PROTEIN INTERACTION VALIDATION

Luisa Montecchi-Palazzi

PhD in Cellular and Molecular Biology-XVIIth cycle

CONTENTS:

GLOSSARY:

Interactome : complete set of protein-protein interactions of a given organism.

Proteome : complete set of protein-encoding Open Reading Frames (ORF) automatically translated from fully sequenced genomes of a given organism.

Protein pattern or motifs : specific cluster of amino acids known to be associated to a specific protein function.

Position Specific Scoring Matrix (PSSM) : probabilistic representation of a protein pattern, using a matrix storing each residue propensity to occur within a sub-sequence matching the pattern.

Regular expression : deterministic representation of a protein pattern taking advantage of symbols to list the allowed residues within a sub-sequence matching the pattern.

Databases acronyms :

GO: Gene Ontology
ELM : Eukaryotic Linear Motif database
SGD : *Saccharomyces* Genome Database
CYGD : Comprehensive Yeast Genome Database
YPD : Yeast Protein Database
MINT: Molecular INTeraction database
BIND : Biomolecular Interaction Network Database
DIP : Database of Interacting Proteins
UniProt: Universal Protein resource

Abbreviations:

ORF : Open Reading Frame
Co-Ip : CO-ImmunoPrecipitation
PSSM : Position Specific Scoring Matrix
BLU : Boheringer Light Units
PepSpot : PEPtides arrays generated by SPOT synthesis

ABSTRACT

Currently, a large community effort focuses on protein interaction data as a mean to explore uncharacterized proteins function, discover new pathways and identify potential drug targets. However, the redundant screenings carried out in the past four years in *Saccharomyces cerevisae* show a very weak overlap and underline the need of protein interaction data validation. Here we propose a new comparative genomic validation approach based on the conservation of binding sequences within orthologs alignments of fifteen closely related yeast species.

Taking the 14-3-3 domains as a study case we explore the binding specificities of their ligand peptides taking advantage of mutagenesis analysis carried out by PepSpot experiments. Using these experimental results we create a prediction tool based on regular expression combined with position specific scoring matrix able to screen the full in *S. cerevisae* proteome and identify putative 14-3-3 domain ligands. The comparative genomic method together with other well established protein interaction validation approaches are benchmarked as filters to increase the accuracy of this prediction. We show that the conservation across several yeast species of 14-3-3 interacting sequences successfully discriminates binding sites from spurious regions matching by chance ligand consensus and increase the prediction accuracy of a four fold.

INTRODUCTION

In the current post-genomic era one of the main issues of molecular biology is the functional characterisation of gene products. Taking advantage of the massive sequence information available we can now list the genes and the proteins encoded by a steadily increasing number of genomes. However, sequence knowledge is just the first step of a “proteome” understanding and is far from enough if we aim at deciphering how cellular networks regulate complex processes such as development or external stimuli responses. Comprehensive protein-protein interaction maps promise to reveal many aspect of regulatory and molecular mechanism underlying cellular function. Moreover, interaction network together with RNA expression data and genomes sequences are the major component whose integration should lead system biologists to re-create comprehensive artificial cellular model in the forthcoming years. To pursuit this aim it is important to establish unified standard to facilitate navigation among different data types and indeed to ensure the reliability of the information provided. Significant efforts are currently done not only to standardize and map related gene RNA and protein sequences (Kersey P *et al.,* 2005) but also to release expression interaction (Brazma *et al.,* 2001) and other proteomic data (Orchard *et al.,* 2004) according to a unified format to the public domain. Our interest focuses on protein-protein interaction, and here we want to briefly summarize the different experimental methods used to determine protein interaction and the issues raised by the recent advance in interaction network exploration in particular the need of data validation procedures.

Overview on protein interaction data

Protein interaction detection methods

Currently we have collected more than a hundred experimental methods described in the literature as tools to detect protein interaction[1]. (See materials and methods section Curation of protein-protein interaction). In table 1 we report only some of the most popular methods and their general

[1] http://cvs.sourceforge.net/viewcvs.py/psidev/psi/mi/controlledVocab

features. Each method detects protein interaction at a specific level of resolution. For instance X-ray crystallography provides an atomic resolution of the interaction interfaces, whereas co-immunoprecipitation (Co-Ip) only detects the proteins participating in an interaction but does not give any clue about the topology of the complex. On the other hand, Co-Ip provides an insight on complex formation within living cells and has been implemented in a high through put scale determining the *in vivo* partners of thousand of proteins. Besides, in order to determine ligand affinity in a straightforward manner others biophysical analysis are required and similarly specific experiments must be performed to assess interaction partners location *in vivo or* to determine the specific residues involved in an interaction.

Class	**Method**	**Analytical Perspective**	**Resolution level**	**Advantages**
Biochemical				
	Co-immunoprecipitation (Co-Ip)	Perspective	Complex	*in vivo*
	ELISA	Analytical	Protein	
	Filter Blot	Analytical	Protein	
	Pull down	Analytical	Protein/residues	
	Co migration in non denaturing gel	Analytical	Complex	*in vivo*
Biophysical				
	X-ray crystallography	Analytical	Atomic	
	NMR	Analytical	Atomic	
	Surface plasmon resonance (Biacore)	Analytical	Protein	kinetics
	Isothermal titration calorimetry	Analytical	Protein	kinetics
	Scintillation proximity assay	Analytical	Protein	kinetics
Protein Complementation				
	Two hybrid	Perspective	Protein	
	FRET	Analytical	Protein	*in vivo* location
	Bacterial two-hybrid	Perspective	Protein	
Array technologies				
	Protein array	Perspective	Protein	
	Pep Spot	Perspective	Residues	
	Phage display	Perspective	Residues	

Table 1 : Main features of some protein interaction detection method. We call perspective the methods that have currently been used to explore uncharacterized protein interaction whereas analytical apply to experimental procedures that investigate known interactions.

Overall there is currently no single interaction detection procedure that can provide alone comprehensive information about an interaction *in vivo*. The only way to collect all possible data about an interaction is through the integration of different experimental results. This is the main scope of protein interaction database such as MINT (Zanzoni *et al.,* 2002), Intact (Hermjakob *et al.,* 2004b), DIP (Xenarios *et al.,* 2002), HPRD (Peri, S. *et al.,* 2004) and BIND (Alfarano *et al.*, 2005) where all experimental evidences supporting a given interaction are reported and uniformly collected in a common repository.

In the last five years two techniques have emerged to explore large numbers of proteins interaction without any prior knowledge of their biology apart from their sequences. Two hybrid (Fields & Song 1989)and Co-Ip are in fact well established experimental methods often used in small scale experiments that have also been adapted to high level of automation for systematic detection of large numbers of interacting proteins.

The double hybrid or two hybrid system is a method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature of the site-specific transcriptional activators GAL 4 (see Figure 1) , which consist of a DNA-binding domain (BD) and a transcriptional activation domain (AD). The DNA-binding domain serves to target the activator to a reporter gene that will be expressed, and the activation domain contacts other proteins of the transcriptional machinery to enable transcription to occur. The two-hybrid system is based on the observation that the two domains of the activator need to be non-covalently brought together by the interaction of any two proteins.

The application of this system requires the expression of two hybrid plasmid encoding the two proteins under study as fusion protein of the activator and DNA binding domain. Several variations of the high throughput implementations of the two hybrid method are reported in the literature and each one has specific features that can deeply influence the reliability of the results. One of these is the two hybrid matrix approach (Uetz *et al.,* 2000) where a collection of haploid yeast strain carrying hybrid plasmids is mated with an equivalent collection of opposite mating type on a microwell plate array.

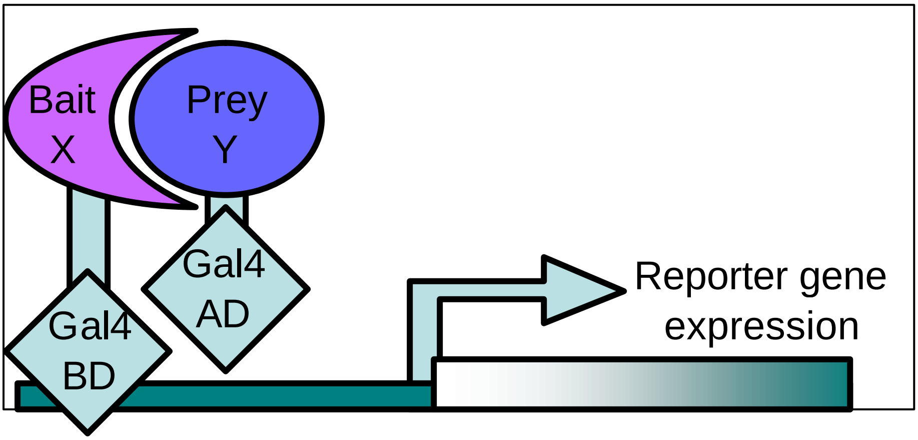


Figure 1 : Two hybrid mechanisms for protein interaction detection. When a polypeptide fused to the binding domain interact with the one fused to the activation domain a reporter gene is expressed.

This approach has the advantage of exploring systematically all possible protein pairs for interaction testing an interaction in both experimental directionalities (A-bait-B-prey and B-bait-A-prey) and thus the investigations are generally limited to hundreds of proteins. However, the interacting protein identity is associated to their position in the mating matrix and re-sequencing of the positives dones is not always performed. On the other hand, in the library pooling approach (Ito *et al.*, 2001) a complex collection of hybrid plasmid are randomly cloned into yeast cells enabling the exploration of thousands of protein interactions but without guarantee of testing all possible combinations. In the library approach the sequence of the positive clones is always determined downstream certifying the proper insertion of the coding sequence in the plasmid. Moreover using the library approach several clones can detect the same interaction, and the number of double hybrids that independently report an interaction is used as an internal reliability measure. Finally when a library of coding sequence fragments are used (Rain *et al.*, 2001), each interaction is redundantly detected and the minimal binding regions required for the interaction can be derived by comparing the clone sequences. In all cases although two hybrid interactions occur *in vivo*, the two protein moieties are artificially co-expressed and brought together in the nucleus to test their ability to interact. However, the two proteins under analysis might never be expressed at the same time or in the same cell compartment in a natural system.

Co-immunoprecipitation instead is a method that can be carried out without altering the natural expression of proteins within cells. In fact Co-Ip relies on an antibody, specific for the protein of interest (generally called bait) or any tag expressed within a fusion protein, used to separate the bait from a protein mixture or a cell lysate and to capture its ligand simultaneously (see Figure 2). The protein partners that bind to the bait protein are retained by the resin are co-eluted, then they are separated by electrophoresis and identified either by immunoblot or by mass spectrometry when the experiment is carried out in a high throughput manner. The exploitation of tags fused to the proteins often implicates transfection of expression plasmids leading, as a consequence, to over-expressing the bait. On the other hand the tags can also be attached to the proteins by genomic integration of the tag coding sequence directly in the genomic copy of the gene under study leaving gene expression under the control of the endogenous promoter. Nevertheless the tags ensure specific retention of the bait in the affinity column and reduce the incidence of contaminant proteins (Ho *et al.*, 2002). For instance engineered tags like the tandem affinity purification (TAP tag) allow a double purification as it encodes a calmodulin-binding peptide and the *S. aureus* protein A separated by a cleavage site (Gavin *et al.*, 2002, Bouwmeester T *et al.*, 2004).

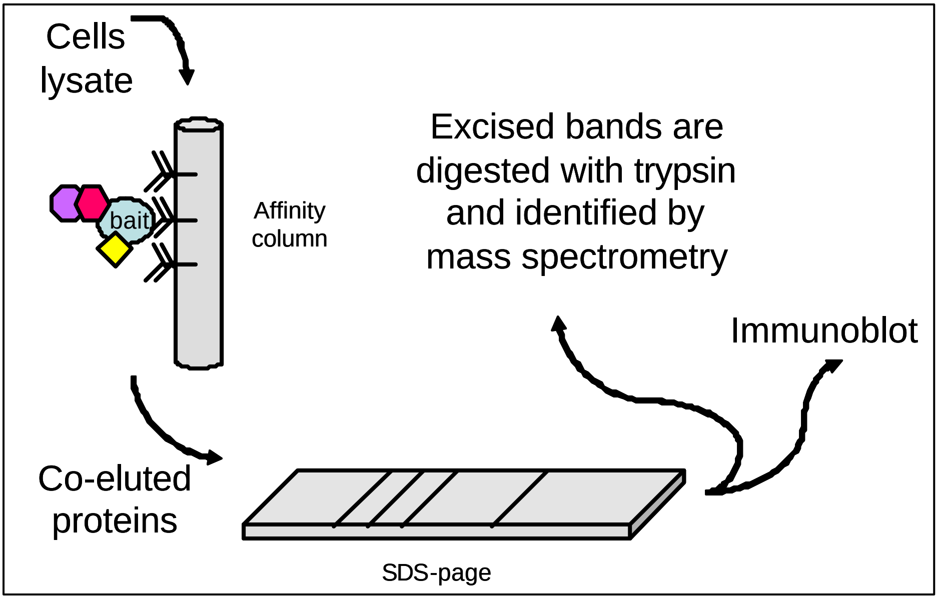

Cells lysate
bait
Affinity column
Excised bands are digested with trypsin and identified by mass spectrometry
Immunoblot
Co-eluted proteins
SDS-page

Figure 2 : Main steps of a co-immunoprecipitation experiment.

Although Co-Ip is generally considered as a method detecting naturally occurring interaction it does not prove that two or more proteins directly bind each other. Co-Ip detects clusters of aggregated proteins without providing any indication of the binary pairs of interacting proteins within a complex. Such cluster are generally deployed according to two alternative schema, the so call "spoke model" where all possible bait-prey pairs are inferred to interact or the "matrix model" where all bait-prey and prey-prey pairs are predicted to be ligand. In fact only complementary experimental results derived from pull down, ELISA, or other *in vitro* methods can determine if two purified proteins directly interact without the intermediation of a scaffold or bridging protein. Even two hybrid results although they consist on binary pairs of interacting proteins do not prove an interaction is direct as whenever a third protein bridge the bait and the prey its presence is not detectable by the system.

Current status of protein interaction exploration

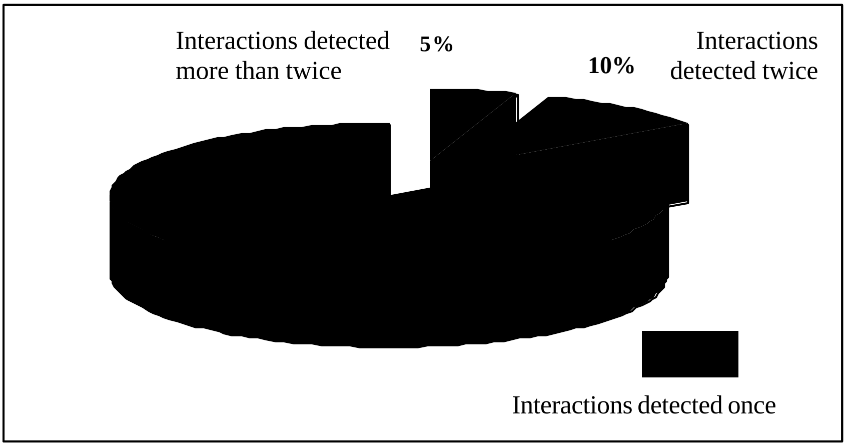
In the past four years many high throughput two hybrid and Co-Ip experiments have been carried out to explore interaction networks of the main model organisms (see Table 2).

Organisms	Complex purified	Binary interaction	Method	Reference
H. pylori		1.465	Two hybrid pooling of fragment library	Rain *et al.,* 2001
S. cerevisae		4.549	Two hybrid pooling approach	Ito *et al.,* 2001
S. cerevisae		1.511	Two hybrid array approach	Uetz *et al.,* 2000
S. cerevisae	589	3.757(*)	Tap tag coimmunoprecipitation	Gavin *et al.,* 2002
S. cerevisae	741	2.583(*)	Flag tag coimmunoprecipitation	Ho *et al.,* 2002
C. elegans		4.624	Two hybrid pooling approach	Li *et al.,* 2004
D. megalonaster		20.676	Two hybrid pooling approach	Giot *et al.,* 2003
H. sapiens	32	1.814(*)	Tap tag coimmunoprecipitation	Bouwmeester T *et al.,* 2004

Table 2 : High throughput interaction detection experiments.

(*) Binary interactions are derived from the co-purified complex only as a set of bait-prey pairs, according to the "spoke model".

In the case of *C. elegans* and *D. megalonaster* most of the known interactions are derived from the high throughput experiments, whereas for human and yeast other small scale experimental results are available in the public interaction databases. Regarding *Homo sapiens* 11.500 unique interactions are spread in the various interaction databases, only 15% have been detected on a single large scale experiment and 26 % of them have been observed in two independent experiments (Cesareni *et al.,* 2005). For *S. cerevisae* around 13.000 unique interactions can be retrieved, more than 90% of those come from high throughput experiments and only 15 % have two or more supporting experiments (see Graph 1).



Graph 1 : Distribution according to their number of supporting experimental evidences of the yeast 13.000 interactions network.

The lack of overlap between the yeast high throughput experiments is striking and has three possible explanations: the various method may not have reached saturation of an estimated 30.000 interactions network (Ito *et al.,* 2002, von Mering *et al.,* 2002), the methods may produce a significant fraction of false positives, some methods may have bias towards certain types of interactions, resulting in the complementarities between the methods.

This observation highlights the necessity of using interaction data with caution and it underlines the need of further experimental investigation to complete our knowledge at least in model organisms of fully covered interaction networks. Moreover it emphasizes the utility of interaction databases as integrative repository of the multiple evidences needed to fully investigate and assess the reliability of protein interactions. Finally the distrust about interaction information obtained by high throughput experiments has stimulated the development of validation procedures to identify interactions that are more likely to be true.

Current approaches to protein interaction validation

Interaction validation by network intersection

The idea of systematically double checking interactions has been first proposed by Tong and co-workers (Tong *et al.,* 2002) that investigate SH3 mediated interactions in yeast both by phage display and by two hybrid. Only 25 % of the interactions are detected by both methods but they show that the intersection is highly significant as the overlap of randomized networks is by far much smaller. Moreover the intersection network they obtain is enriched in interaction derived from the literature over threefold compared to the single method networks. Thus the authors conclude that considering intersection network is a simple but powerful mean to exclude the false positive results of each experimental screening.

Similar conclusions come out of the systematic comparison of all large scale yeast data sets benchmarked against a trusted reference set of manually curated interactions derived from the literature (von Mering *et al.,* 2002). This study confirms that the highest accuracy in recovering known interactions is achieved by deriving intersection networks where every interaction is supported by any pair of evidence and that the overlap of high throughput data is twenty times larger than expected by change. Furthermore the overlap networks mainly consist of interaction in which both partners have the same functional annotation and cellular localisation.

Accordingly in the MINT database a general confidence value is assigned to every interaction based on the following criteria :
Confidence level 1: interaction detected by a single experiment
Confidence level 2: interaction detected by two or more experiments reported in the same publication.

Confidence level 3: interaction detected by two or more experiments from independent publications.

Confidence level 4: interaction detected by two or more independent experiments and at least one of those is an *in vivo* Co-Ip.

Confidence level 5: interaction detected by two or more independent evidences and at least one of those is an *in vivo* Co-Ip experiment carried out using antibodies against endogenous proteins.

This simple tool provides a straightforward estimation of the reliability of each the interaction stored in the database.

Interaction validation by co-annotation

This validation method has been currently used only in *S. cerevisae* not only because its interaction network is the most covered but also because yeast as a model organisms has a very well characterized proteome. More than 4.500 of its predicted 6.200 ORFs have been verified and functionally investigated. The three main repositories dedicated to *S. cerevisae* are Comprehensive Yeast Genome Database (CYGD) (Guldener *et al.*, 2005), *Saccharomyces* Genome Database (SGD) (Balakrishnan *et al.*, 2005) and Yeast Protein Database (YPD) (BIOBASE Wolfenbuttel, Germany) and they all provide functional annotation about gene products either using Gene Ontology cross-reference (See Material and methods section about Gene Ontology) or adopting internal classification systems. Generally a characterized gene product have annotation concerning its subcellular location, its molecular function (such as enzymatic activity, or DNA binding ability) and its cellular-role that refers to its participation in broad cellular process (for instance metabolisms or stress response). Trivially proteins that interact should share the same subcellular compartment and are likely to participate in the same cellular process whose mechanism may rely on their binding. However binding partners can have very different molecular function as for instance a regulatory interaction can easily involve a protein kinase and a transcription factor. Thus co-location and the sharing of a common cellular-role can be used as supporting evidence of a true positive interaction whereas having the same molecular function cannot be employed as a validation criterion.

The observation that in trusted interactions the protein partners share both their functional annotation in terms of cellular processes and cellular compartment (von Mering *et al.* 2002) can be used to quantitatively estimate the accuracy of experimental data sets (Sprinzak *et al.,* 2003). In this analysis the rate of true positives is measured for high throughput and for small scale data clustered according to the method used for their detection. The results provide a possible ranking of the various experimental methods and confirm once again that intersection networks are the one with the highest rate of true positives interactions. According to this study affinity chromatography and Co-Ip are the more reliable methods whereas approximately 50% of the two hybrid large screenings are believed to be false positives.

Interaction validation by co-expression

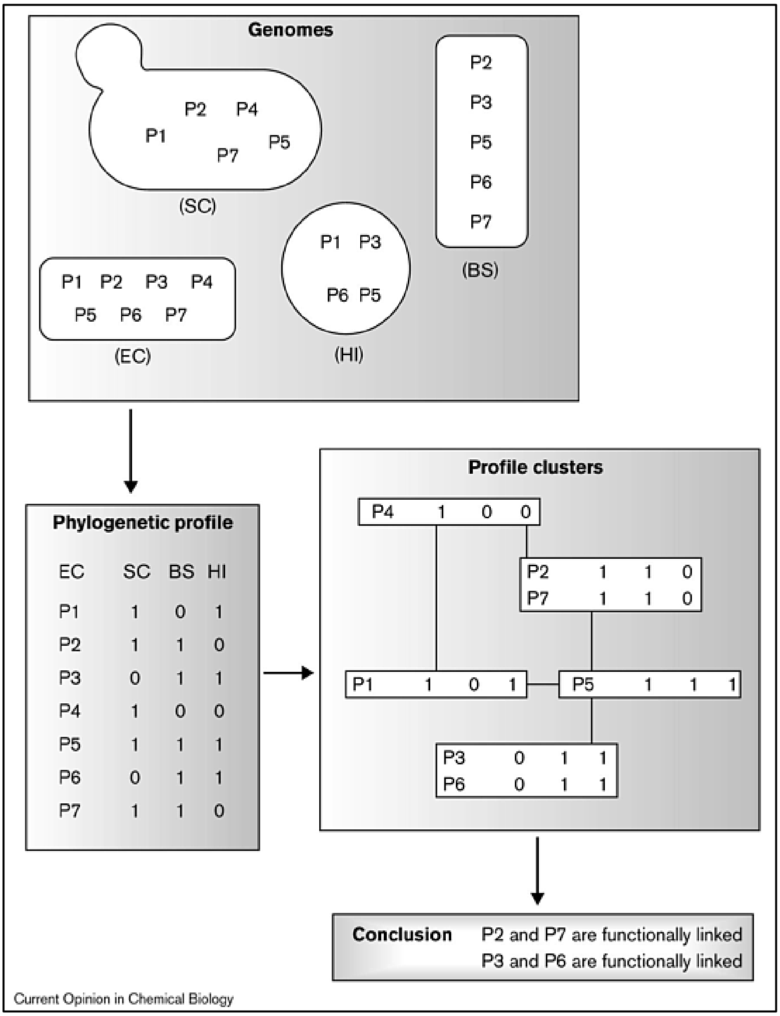
The basic hypothesis underlying such approach is that genes that are co-regulated and co-expressed are likely to be involved in related biological process and thus to be interaction partner. Yeast remain the ideal model organism to carry out this analysis because together with the interaction data large scale RNA expression data are available (Giaever *et al.,* 2002, Kemmeren *et al.,* 2002). The first evidences supporting this well accepted idea are provided identifying clusters of genes having the same expression pattern under different conditions and subsequently counting the protein interactions occurring within or across expression clusters (Ge *et al.,* 2001). In this work it is shown that the number of interactions within any given expression cluster are higher than expect when the same calculation is done on random networks. This confirms that co-expressed proteins more frequently interact among each other than with protein having different expression profiles.

Further studies from the same group focuses on highly connected protein the so called "hub" of the interaction network (Han *et al.,* 2004). By measuring the correlation in RNA level between each hub and its partners two distinct populations are identified: the so called "party hub" that have high expression correlation with their partners (corresponding to the previous intra-cluster interaction) and the "date hub" that show lower expression correlation. Such distinction leads to a model for interaction network where “party hubs” define the core of subnetwork modules whereas “date hub” ensure interconnection between the various modules. In fact the “party hub” partners have very homogenous annotation in terms of process and subcellular location and often correspond to functional complexes or known pathways. “Date hubs” do not show such homogeneity but interestingly their central role as network organizer is confirmed by the lethality of their gene deletion observed in genetic perturbation screenings.

Other independent works support the statement that among the various types of protein interaction mainly protein functional complexes, such as RNA polymerase or the ribosome, show a strong correlation of RNA expression levels (Jansen *et al.*, 2002). Thus certainly co-expression have an average relation with protein interaction data and a strong expression correlation is a valuable *in vivo* supporting evidence that two or more proteins are likely to interact. However the problem is this approach do not help in identifying false positives interactions while lower correlation among partners could also be biologically meaningful (see above “date hubs”). Nevertheless a validation approach based on co-expression is currently available at DIP database. Expression data are not used to validate single interactions but to measure the overall reliability of a given interaction screening (Deane *et al.*, 2002). This method calculate expression level Euclidean distance for all protein pairs in dataset under analysis and compare the distribution of such distances with the corresponding values derived from a trusted reference set of interaction and from a completely random network. The distance distribution calculated for two-hybrid screenings in yeast being equidistant from reference and random network distribution plots leads to the authors conclusion that overall 50% of the interactions are false positives.

Interaction validation by genome comparison

Genomes comparison can also lead to protein interaction validation. Many methods look for the evolutionary relics of protein interaction across fully sequenced genomes to infer functional or physical interaction. For instance one method, so called the Phylogenetic Profile method (see Figure 3), implies that proteins that are consistently present or absent in different proteomes sets are likely to have a functional relationship (Pellegrini *et al.*, 1999). A second one the Rosetta Stone method searches proteomes looking for proteins that are covalently joined in a single peptide chain and interpret this fusion event as evidence that these two proteins interact either physically or functionally in other organisms (Marcotte *et al.*, 1999, Enright *et al.*, 1999). A third one the Gene Order method, look at relative position of genes on various prokaryotes genomes and infer that genes submitted to an evolution pressure to remain close are likely to share the same function (Dandekar *et al.*, 1999).



Phylogenetic profile

EC	SC	BS	HI
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0

Figure 3 : Phylogenetic Profile method to detect protein functional interaction (Pellegrini 2001).

The protein-protein associations detected with these three methods are collected in the STRING database (von Mering *et al.*, 2005) together with the experimental protein interactions and data derived from functional pathways. Data cross-comparison show that the protein relationship detected by genomic methods often correspond either to physical interactions, or to the participation in a same metabolic pathway or cellular process. Thus genomic methods can provide supporting evidence for protein interaction validation and are powerful tools for functional characterisation of newly discovered sequences. The three methods predict protein function with have an average 80% confidence in prokaryotes model organisms (Huynen *et al.*, 2003).

Aside from these methods which explore genome organisation other focuses on conservation at the amino acid sequence level. One instance is the "correlated mutation" method that exploits parallel sequence variations in multiple alignments of orthologous sequences to infer interaction (Pazos & Valencia 2002). The underlying rationale is that pair of residues that are part of two interacting surfaces tend to co-evolve with changes in one protein being remedied by compensatory mutations in the partner protein. When there is a good species coverage for any protein pair under study application if this method by mapping residues involve in an interaction interface can also provide topological information about the interaction complex.

PART 1: THE EXPLOTATION OF THE PEPSPOT TECHNOLOGY TO INVESTIGATE PEPTIDE BINDING PREFERENCES

PepSpot, or peptides arrays synthesized on membrane by SPOT synthesis (Frank 1992, Reineke *et al.,* 2001), is a powerful method to investigate molecular recognition events and has been heavily used in our laboratory to determine interaction between short peptides (8-25 residues) and their binding domains. The PepSpot method is not only suitable for protein interaction detection, but it gives direct access to the sequences responsible for the interactions enabling further binding specificities analysis and development of prediction tools.

PepSpot as a protein interaction detection method

PepSpot technology

The PepSpot method provides a fast procedure to generate a large number of peptides and screen for their biological properties. The method applications (see Figure 1) include: the identification of peptides that promote immunological activity (epitope mapping), the characterization of enzyme targets (enzyme-substrate screening) and the detection of protein binding peptides (protein binding screening).

Focusing on the PepSpot *in vitro* protein binding screening, the experimental procedure (Kramer & Schneider-Mergener, 1998) can be summarized in four main steps:

1. Preparation of the membrane before the peptide synthesis. The Cellulose-Amino-hydoxyPropyl Eter (CAPE) membrane is activated by treatments with a number of reagents ensuring the covalent binding of the C-terminus of each peptide.
2. Peptides are automatically synthesized on the cellulose membrane using a SPOT synthesizer under positional control of LISA software (Jerini AG, Berlin Germany) guaranteeing that a specific sequence including residue modifications is synthesized at any given position.

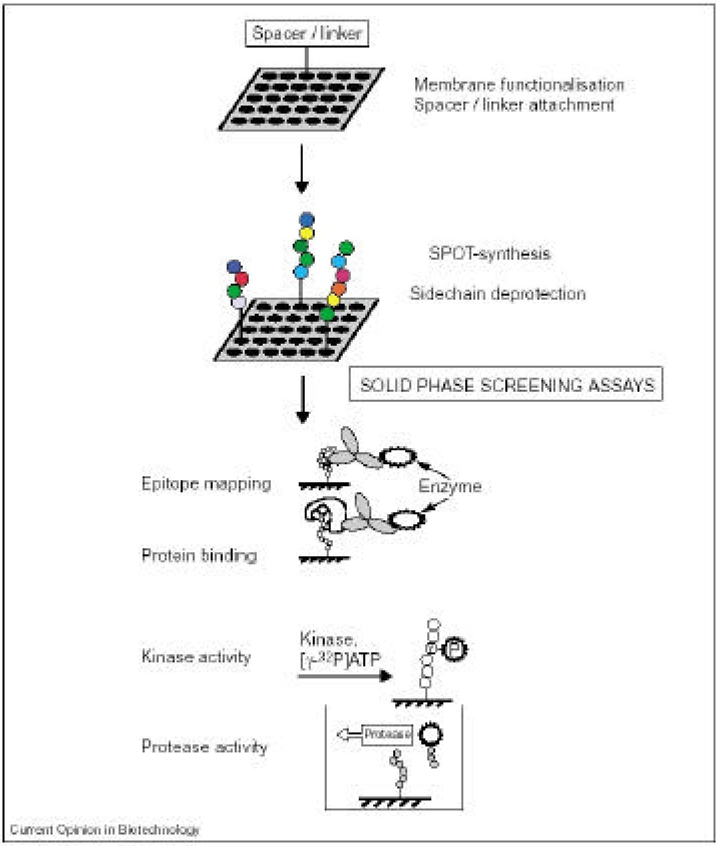


Figure 1: Different application of the SPOT-synthesis peptide arrays (Reineke *et al.*, 2001).

3. Incubation of the membrane with a probe protein previously expressed and purified.
4. Detection of the spot retaining the probe after membrane washes, using anti-probe antibodies. The detection of a peptide-bound protein can be carried out using a chemoluminescence substrate providing a quantitative intensity signal in BLU (Boheringer Light Units), associated to any spot.

The PepSpot method has the great advantage of reproducing synthetically any naturally occurring sequence and any kind of post translation modification on it. On the other hand, the extent of the interaction space that can be explored with this method is clearly limited by the peptide length and by the number of peptides (around 10.000) that can be reasonably synthesized given the current technology. However, a fairly large part of protein interaction relies on small domains (SH2, SH3, PTB, WW, EH, PDZ, GYF, VHS, WD40, 14-3-3, FHA etc.) accommodating in their binding pocket short peptides of the partner protein in extended conformation (Pawson & Nash, 2003). Thus, in a good number of cases it is reasonable to model protein-protein interaction to peptide-domain interaction and consequently use the PepSpot method to identify natural peptides with the potential for binding a given domain. Clearly, the inference of *in vivo* interactions from PepSpot results could turn out to be incorrect for several reasons. First, the domain recognition determinants on a protein surface may be dispersed discontinuously on the sequence and may not be represented by any linear peptide. Alternatively, a potentially binding peptide could be buried inside the folded protein and therefore inaccessible to interaction partner. Finally, the two inferred partners might never coexist *in vivo* because they are located in different cellular compartments or expressed in different tissues or at different times during an organism development.

Proteomic screening of SH3 mediated interactions with PepSpot method

To assess the feasibility of a proteomic screening approach by the PepSpot method, eight *S. Cerevisae* proteins containing SH3 domains (RSV167, YFR024C, YSC84, ABP1, MYO5, SHO1, BOI1 and BOI2) were chosen and tested against two set of peptides (Landgraft *et al.*, 2003). The peptide sets are generated searching all the peptides of the *S. Cerevisae* proteome that match the two well established binding *consensi*, the class 1 ([RK]xxPxxP) and the class 2 (PxxPxR) motifs. Each SH3 domain derived from the above mentioned proteins is probed against approximately 1,500 peptides synthesized at high density on cellulose membranes (see Figure 2). 211 peptides resulted to bind one or more of the probe SH3 domains with an intensity higher than a threshold of 20,000 Boheringer Light Units (BLUs), corresponding to a dissociation constant of approximately 10^{-6} M. The derived network counts 180 inferred interactions (see Figure 3) among 111 proteins (some proteins contain more than one high affinity peptide).

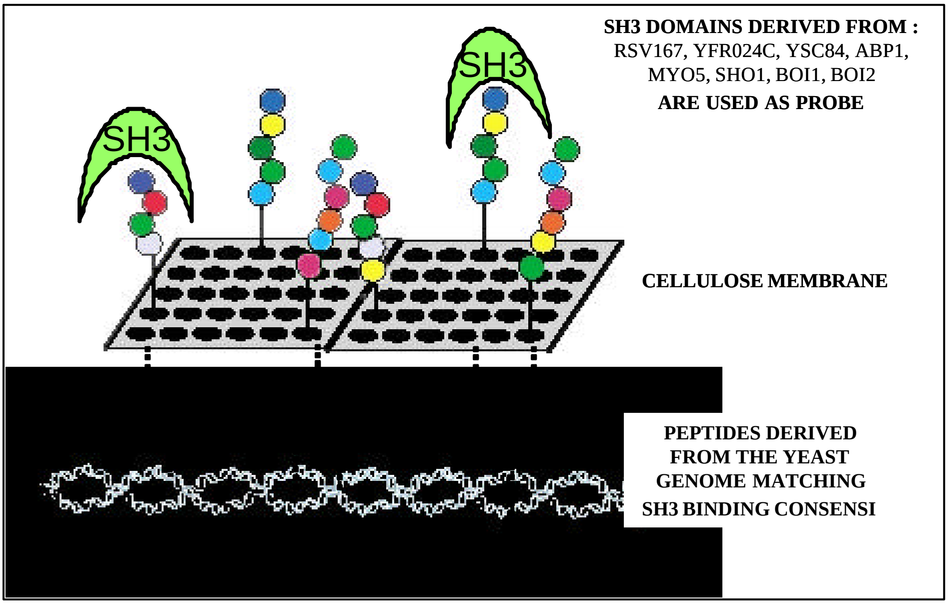


Figure 2 : Strategy to explore all SH3 mediated interaction in *S. cerevisae*.

As shown in figure 3, the SH3 domains of RSV167, YFR024C and YSC84 have overlapping specificity and thus promiscuous binding partners, whereas ABP1, SHO1, BOI1 and MYO5 have almost exclusive interactors. In the case of BOI2, the fact that it binds only 3 interactors with high affinity, confirms that its binding specificity is divergent from the standard motifs. This underlines the danger of extending a binding motif to all members of a protein family without experimental validation.

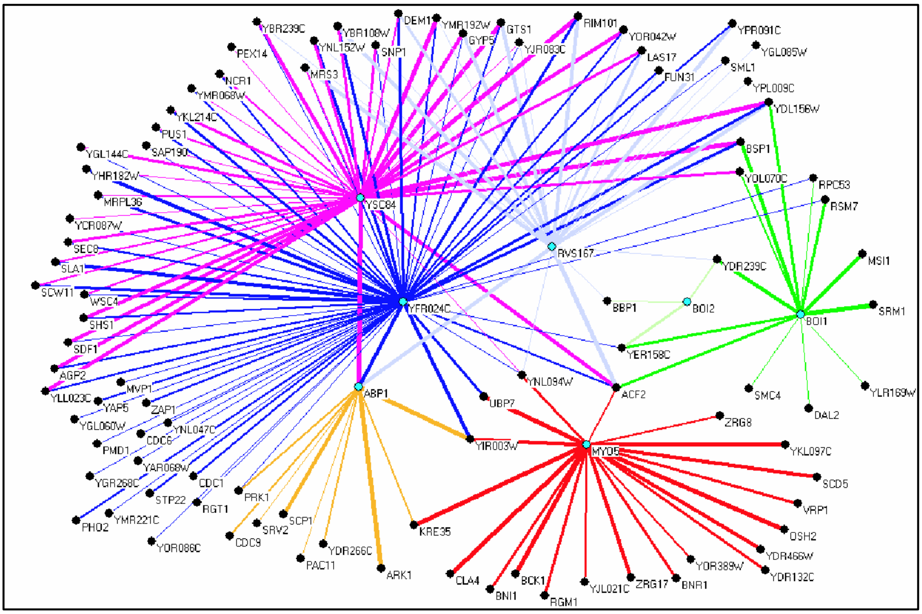


Figure 3: Protein interaction network inferred from PepSpot results. Nodes in cyan represent the SH3 containing proteins used as probe in the screening. The edge thicknesses are proportional to the BLU intensity of the corresponding interaction.

PepSpot network comparison

The PepSpot approach when compared with two hybrid and complex purification (see Table 1), offers the advantage of providing direct information about the sequences responsible for interaction and an estimate of the dissociation constant, thus complementing the information obtained by more direct *in vivo* experiments. Moreover, the lack of a substantial overlap between the results of the high throughput projects meant to cover the entire interactome of *Saccharomyces cerevisae*, emphasizes the importance of confirming any interaction by different methods (von Mering *et al.*, 2002).

Method	Cost of the screening	Interaction detected	Binding sequence identification	Affinity measure	Other bias
PepSpot	Large screen (5.000 peptides on a single membrane) fast and relatively low cost.	*In vitro* binary interaction.	The domain recognition target peptides are clearly identified	Signal intensity inversely correlated with the dissociation constant	Restricted to small peptide-domain interaction
Two hybrid	Variable cost depending on the automation level and on the extent of the screen.	Binary interactions detected *in vivo* in a non physiological context.	Target identification requires the use of a fragment library and the comparison of several overlapping clones.	In low throughput experiments beta-galactosidase assays provide estimate of the interaction strength.	Sticky preys Self activating baits
Complex CoIp	High cost for specialized instrumentation.	*In vivo* complexes are detected.	None by standard approach.	None by standard approach.	Proteins are often overexpresse d. Low affinity interactions are missed.

Table 1: Comparison of the properties of three protein interaction methods.

In order to compare networks resulting from the different methods and therefore evaluate the PepSpot method performance, we computed the intersection of all major yeast interaction dataset stored in the MINT database (see Table 2, where the PepSpot SH3 dataset is reported as Landgraft *et al.*).

For each dataset we consider only non redundant interactions (if an interaction is detected twice as A-B and also as B-A we only count it once); by measuring the extent of the intersection network, we also pay attention to count interactions regardless of their directionality and to avoid double scoring of homodimers.

The intersection networks derived by combining pairwise the yeast two hybrid screens (Ito *et al.,* 2001 and Uetz *et al., 2000*) and the high throughput analysis of complexes (Gavin *et al.,* 2002 and Ho *et al.,* 2002), are consistent with cross comparison previously reported in the literature (Ito *et al.,* 2002 and Bader & Hogue, 2002). Besides, with these well known sets, we compare results of the yeast SH3 containing protein two hybrid experiment performed by Tong and co-workers, that contain all the SH3 domains considered in our studies (Tong *et al.,* 2002). Finally, we benchmark every single set against small scale experiments stored in MINT data base and against an automatically generated network including all known yeast interactions, excluding those from the set under analysis (row named "all" in Table 2). On this latter intersection network we evaluate the percentage of interactions and the percentage of interacting proteins shared by a single set with any other one. Obviously, these two numbers are related, because two networks can hardly have a common interaction if the two protein partners are not present in both sets. Thus, looking at the ratio of the overlapping interaction per common protein, we have a better estimate of the ability to recover an interaction of each set.

From these numbers it is easy to see that the network inferred from PepSpot results has a similar performance in comparison to the two hybrid screen carried out by Tong *et al.,*. Finally, although the overall coverage in terms of number of interactions explored is low, PepSpot method seams to have the highest capacity to recover interactions once the proteins of interest are submitted to a binding experiment.

	Dataset size	**Intersect with Ito *et al.***	**Intersect with Uetz *et al.***	**Intersect with Tong *et al.***	**Intersect with Gavin *et al.***	**Intersect with Ho *et al.***	**Intersect with Landgraft *et al.***
Ito *et al.*							
interactions	4443						
proteins	3236						
Uetz *et al.*							
interactions	942	188					
proteins	996	278					
Tong *et al.*							
interactions	231	5	2				
proteins	144	9	3				
Gavin *et al.* (*)							
interactions	3757	62	51	5			
proteins	1471	102	82	6			
Ho *et al.* (*)							
interactions	2583	56	50	5	152		
proteins	1362	95	81	7	189		
Landgraft *et al.*							
interactions	180	2	1	32	3	5	
proteins	111	3	2	25	4	8	
Small scale data							
interactions	956	29	37	30	56	34	20
proteins	794	49	60	32	76	56	20
All ($)							
interactions		276	266	53	272	238	41
proteins		391	376	42	324	302	30
% Overlapping interactions		**6.21**	**28.24**	**22.94**	**7.24**	**9.21**	**22.78**
% Overlapping proteins		**12.08**	**37.75**	**29.17**	**22.03**	**22.17**	**27.03**
% Ratio (ovrlp. Int. / ovrlp. Prot.)		**51.41**	**74.80**	**78.66**	**32.87**	**41.55**	**84.28**

Table 2: Interaction data set cross comparison.

(*) Binary interactions are derived from the co-purified complex only as a set of bait-prey pairs, according to the "spoke model".

($)The "All" set differs for every single dataset and it is generated by combining all known yeast interactions and excluding only the set under analysis.

SH3 mediated network comparison

Focusing on the comparison of SH3 domain connectivity, we can group interaction networks according to detection method (see Table 3) and generate SH3 mediated subnetworks, by assuming that SH3 containing proteins interact exclusively via this domain.

Dataset	**Dataset size**	**Number of SH3 containing proteins**	**Inferred SH3 mediated network**	**Average interactions per protein in the original dataset**	**Average interactions per protein in the inferred SH3 mediated network**	**Average interactions per SH3 containing protein in the inferred network**
Complex CoIp Gavin + Ho						
interactions	6188		94	2.78	1.04	4.70
proteins	2225	20	90			
Two hybrid Ito + Uetz						
interactions	5197		141	1.47	0.94	8.29
proteins	3535	17	150			
Any method Small scale data						
interactions	956		109	1.20	1.28	6.06
proteins	794	18	85			
Two hybrid Tong *et al*.						
interactions	231		231	1.60	1.60	8.56
proteins	144	27	144			
PepSpot Landgraft *et al*.						
interactions	180		180	1.62	1.62	22.50
proteins	111	8	111			

Table 3 : Network comparison of inferred SH3 mediated interactions.

In each dataset we identify all the proteins containing an SH3 domain, and retrieved their interaction partners. The subnetworks derived from the first three sets are inferred to be mediated by SH3 whereas the last two set are experimentally determined using exclusively the SH3 domains as probe.

Then we calculate the average connectivity of all proteins in each subnetwork and compare it with the corresponding value in the original network. We expect these two values to be similar for each dataset. This would mean that the analysis of SH3 domain containing proteins do not affect each method ability to detect interaction. Finally we calculate the average number of interactions per SH3 domain inferred or detected using each experimental method. These numbers allow the comparison of how many partners are identified on average for an SH3 domain by the single methods.

Although this is an inference, since only a fraction of the interactions in the fourth column are mediated directly by SH3 domains, it is nevertheless surprising to see how far SH3 mediated interactions seem to be under-detected by complex purification approach. The average number of interactions per protein drops by more than 50% when the SH3 subnetwork is considered and the number of inferred interactions per SH3 is also much lower than in any other method. This is consistent with the fact that SH3 mediated interactions are on average weak interactions, less likely to be detected by coimmunoprecipitation (assay requiring several washing steps), than by solid phase peptide arrays or two hybrid assay. It is also evident that the testing of all possible partner peptides, regardless of their physiological co-occurrence with the probe domain, leads to a very high number of PepSpot interactions per SH3, and that some of those are likely to be false positive.

Although the PepSpot proteomic approach we have presented is an *in vitro* strategy limited to interactions in which one of the partners can be reduced to a relatively short peptide, it has a series of interesting features. First, genomic sequence information can be fully exploited in the array format to equally display a high number of possible partners. Second, the experimental output provides topological information and includes an estimate of the kinetic dissociation constant. Third, interactions that depend on peptide modification can be easily studied. Finally, the identified target peptide can be used as a template to develop tighter binding competitors.

In fact the major advantage of the PepSpot method is the easy modulation of the protein binding screening to different level of granularity (see Figure ?) :

1. Proteome scanning (used for yeast SH3 mediated interaction screening). All the peptides of a proteome that match the binding motif

are synthesized and tested for their ability to undergo interaction with probe proteins.

2. Ligand scanning. All the overlapping peptides derived from a known ligand are tested for their ability to interact with the ligand partner to identify the best binding region.

3. Peptide mutagenesis scanning. All possible mutations of a single binding peptide are synthesized, replacing in each position the wild type amino acid with all other possible residues, leaving the rest of the sequence unchanged. The results of this PepSpot variation give indication of position specific information about the residues that are favourable to binding. As we will explain in the following section, such data can be used to estimate the binding ability of any query peptide, and thus predict potential interaction partners for the probe protein.

14-3-3 partner prediction based on PepSpot mutagenesis

14-3-3 containing proteins

In the remaining parts of this work our study case are 14-3-3 proteins in yeast. The 14-3-3 is a family of highly conserved proteins found in almost all eukaryotic organisms. In mammalia seven 14-3-3 encoding genes were found and their study revealed that they are functionally related to crucial cellular processes, such as signal transduction, metabolism, cell cycle, apoptosis and malignant transformation (Mackintosh, 2004). 14-3-3 regulates the function of their partners by binding to them. This may lead to alteration of the ligand catalytic activity, to interference with molecular interactions with other partners, or to regulation of the bound protein subcellular localisation (Yaffe & Elia, 2001). Interaction screenings in mammalia show that 14-3-3 isoforms have hundreds of partners belonging to very different functional classes (Rubio *et al.,* 2004 Jin *et al.,* 2004 Aitken *et al.,* 2002*).*

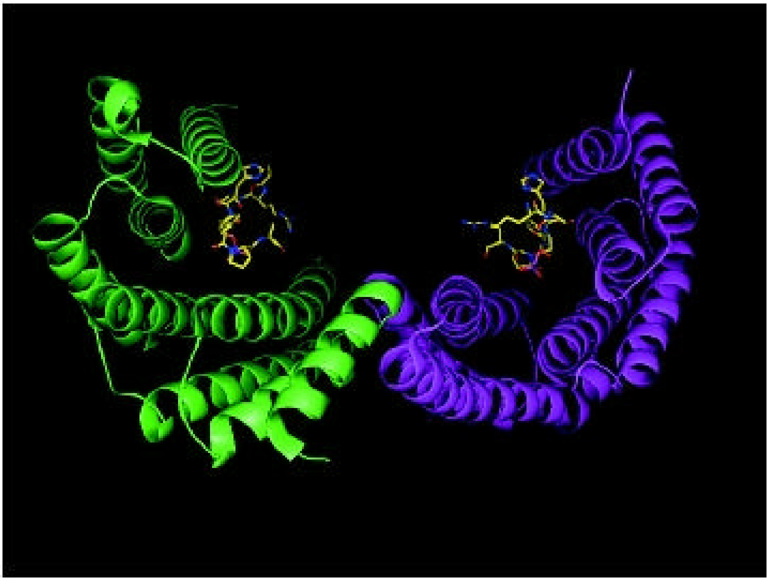


Figure 4 : 14-3-3 pocket showing a phosphopetide ligand bound to each monomeric subunit.

Typically 14-3-3 domain proteins recognize phosphopeptides where serine or threonine is phosphorylated. The X-ray analysis of 2 mammalian isoforms (Liu *et al.,* 1995 Xiao *et al.,* 1995) reveal the dimeric structure of 14-3-3 containing proteins forming a large cup shaped pocket able to bind two peptides in an extended conformation (see Figure 4). The screening of degenerated phosphoserine peptide libraries against human and yeast 14-3-3 (Yaffe *et al.,* 1997) lead to the identification of two 14-3-3 binding motifs (RSxpSxP and RxxxpSxP), while no specie nor isoform specificity was observed. However, a discrete number of interactors have significantly divergent binding sites.

Despite the large number of studies concerning 14-3-3 proteins in higher eukaryotes, little is known about their function in *S. cerevisae*. In yeast, two 14-3-3 proteins have been characterized, encoded by the BMH1 and BMH2 genes. They are more than 90% and 60% identical respectively to the mammalian epsilon isoform. The single genetic disruption of BMH1 or BMH2 has little effect on cellular growth, while the double mutant is not viable (Bruckmann *et al.*, 2004). These two proteins seem to be functionally redundant but essential for normal yeast growth.

PepSpot mutagenesis of 14-3-3 binding peptides

To investigate BMH1 and BMH2 binding specificity, we choose to mutagenize two peptides: IPAWLpSLPS and SRIPFpSERK, known for binding 14-3-3 protein, which match only partially the motifs (RSxpSxP and RxxxpSxP).

At a first glance, by analysing the membranes (see Figure 5) we can see that BHM1 and BMH2 have almost identical recognition specificity. According with the binding motifs proposed by Yaffe, positive charges (R, K or H residues) at the peptide N-terminus increase the binding, while negative charges in the same position considerably reduce it. In all membranes position -1 and +1 are the more restrictive in terms of allowed residues [1]. Position +2 has different preferences depending on the mutagenised peptide. While in the case of IPAWLpSLPS position +2 is quite tolerant, with a slight preference for the wild type proline, in the case of SRIPFpSERK only the wild type arginine is allowed. This points out the influence of the sequence context surrounding the single residues. Although according to Yaffe and collaborators proline in position +2 should enhance the binding, in the context of the SRIPFpSERK sequence it diminishes it.

[1] *In our numbering the phosphorylated serine positions, is positive 0, while residues at the C-terminus or the N-terminus side have positive or negative numbers respectively.*

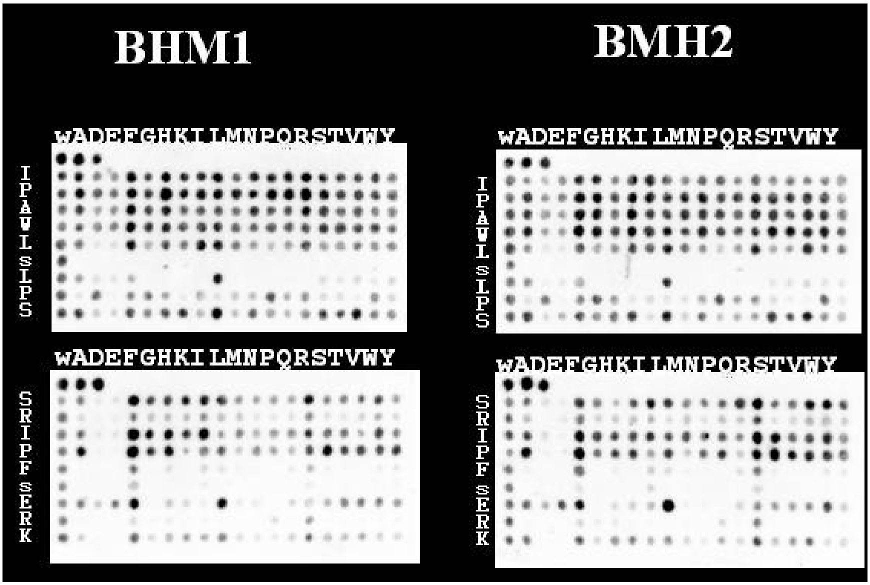


Figure 5 : 14-3-3 binding peptides mutagenesis. In the first column of each membrane the wild type peptide is spotted for every row. In the serine column, amino acids are replaced with unphosphorylated serine: this explains the lack of signal on the central phosphoserine row. The spot triplets in the left upper corners are controls of the colorimetric detection.

Implementation of a prediction tool based on PepSpot mutagenesis

Although there is clear evidence that the results of mutagenesis experiments are not independent from the template peptide, we use the intensity of the colorimetric reaction measured in Boehringer Light Unit (BLU) to build a matrix describing each amino acid preference at each position within a generic phosphoserine centred motif (xxxxxpSxxx). To calculate each element of this matrix we use the following formula :

$$Aff(pos,aa)=BLU(pos,aa)/\ (\Sigma BLU(pos)+ BLUwildtype(pos))$$

These figures are related to the preference for each specific amino acids relative to all other possible amino acids at any same position. The results obtained from the different mutagenesis experiments are combined by calculating an average matrix with the intent of alleviating the bias of any specific peptide context. Finally, the 14-3-3 position specific matrix that we obtain (see Table 3) stores relative preference values, ranging from 18% to 0.06% (proline in position +1). The value for the phosphoserine in position 0 is arbitrarily set to 100%.

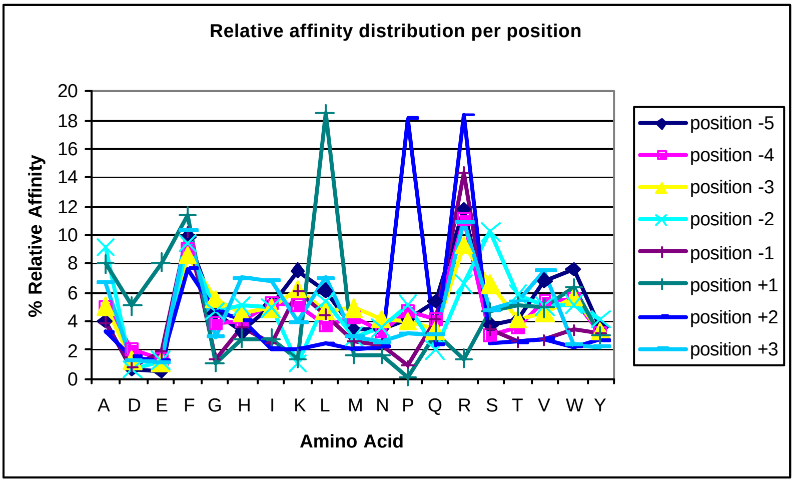
	-5	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
A	3.99	5.09	5.08	9.2	3.86	0	7.97	3.29	6.69
D	0.65	1.97	1.3	0.7	0.84	0	5.09	1.49	1.34
E	0.55	1.37	1.12	1.26	1.87	0	8.16	1.33	1.2
F	9.93	9.07	8.6	9.47	10.2	0	11.3	7.68	10.3
G	4.55	3.87	5.5	4.75	1.47	0	1.11	4.76	2.93
H	3.31	3.97	4.62	5.08	3.49	0	2.84	4.12	7.03
I	5.2	5.26	4.92	4.92	2.43	0	2.8	2.16	6.8
K	7.51	5.2	6.21	1.06	6.11	0	1.38	2.11	4.02
L	6.11	3.71	4.68	5.38	4.52	0	18.4	2.44	7.11
M	3.42	4.25	4.91	2.96	2.64	0	1.62	2.07	2.92
N	3.39	3.29	4.15	3.48	2.3	0	1.69	2.26	2.58
P	4.15	4.66	4	5.2	0.91	0	0.06	18.1	3.23
Q	5.4	4.17	3.43	1.97	4.16	0	3.18	2.3	3.05
R	11.8	11.0	9.33	6.53	14.4	0	1.47	18.4	10.9
S	3.77	3.09	6.54	10.2	3.44	100	4.69	2.51	4.78
T	4.12	3.6	4.08	5.84	2.58	0	5.01	2.55	5.4
V	6.79	5.4	4.57	5.03	2.73	0	4.9	2.78	7.57
W	7.68	5.49	5.7	5.18	3.44	0	6.37	2.22	2.33
Y	3.66	3.23	3.31	4.09	3.19	0	3.02	2.66	2.26

Table 3 : 14-3-3 position specific matrix. Higher scores are highlighted in green, lower in red.

In order to implement a prediction tool based on this data that could rank any query peptide for its ability to bind 14-3-3 yeast domains, we follow a combined approach. Combining the strategies of two major motif public resources, SCANSITE (Obenauer *et al.*, 2003) and ELM (Punterwoll *et al.*, 2003), we used the 14-3-3 preference matrix as a position specific scoring matrix (PSSM), but we also to derived a regular expression that could filter out all the peptides containing ill tolerated residues at any specific position. SCANSITE and ELM services will be presented in detail in the next section while a general description of PSSM and regular expressions is given in the materials and methods (see ‘Protein functional patterns’ section). By analysing the score distribution in our matrix (see Graph 1), we can arbitrarily distinguish three major classes. The "forbidden" residues of each position have a score below 2%, the "neutral" residues range from 2% to 10% and the "preferred" amino acids score above 10%. By using the scores below 2%, we designed a regular expression representing the exclusion of specific residues at each position (coded by the symbol [^]) and setting the requirement of a serine at position 0. The resulting regular expression is:

[^DE][^DE][^DE][^DEKQ][^DEPG]**S**[^RPKGINMQ][^DE][^DE]
Position -5 -4 -3 -2 -1 **0** +1 +2 +3

We first apply this regular expression and then we rank the matching peptides by applying the position specific scoring matrix.



Graph 1 : Analysis of the position specific matrix values. Note that position 0 is not reported because it has to be a phosphoserine.

Our approach has the twofold advantage of overcoming the strict rules imposed by a stringent regular expression designed to identify the best binding motif, by setting as mandatory the preferred residues, while at the same time restricting the number of false hits by using a "softer" filtering regular expression. This flexibility suits the case of 14-3-3 peptide recognition, since, for instance, a positive charge is required at the amino terminus of the ligand peptide, but apparently there is no absolute requirement at any specific position. Thus, by using a PSSM, the best ranking ligand peptides will result from any combination of high scoring residues alternated with tolerated residues. Whereas peptides having residues that in the mutagenesis experiments are shown to compromise drastically the binding are discarded by the application of the regular expression.

Benchmark of the prediction tool versus experimental results

In order to assess the validity of our tool to infer 14-3-3 target peptides, we perform a blind test with new experimental results. For this purpose we carry out a PepSpot scanning on the fraction of the yeast proteome experimentally proven to contain phospho-residues. In a recent work, Ficarro and co-workers describe a methodology to characterize phosphoproteins from cell lysate, using immobilized metal affinity chromatography (IMAC) to purify phosphopeptides and nanoflow HPLC/electrospray ionization mass spectrometry to identify the phosphopeptides (Ficarro *et al.*, 2002). They report the sequence of 216 yeast peptides containing a phosphorylated serine, threonine or tyrosine: 60 of these are phosphorylated in the single position, 145 in two positions and 11 in three. Using these results, we therefore synthesise 287 phosphopeptides on cellulose membrane and assay their ability to bind BMH1 and BMH2 (see Figure 6). Peptides that are found to have multiple phosphorylation sites *in vivo* are synthesise both as single phosphorylated peptides, as well as peptides containing phosphorylated residues at two and three positions.

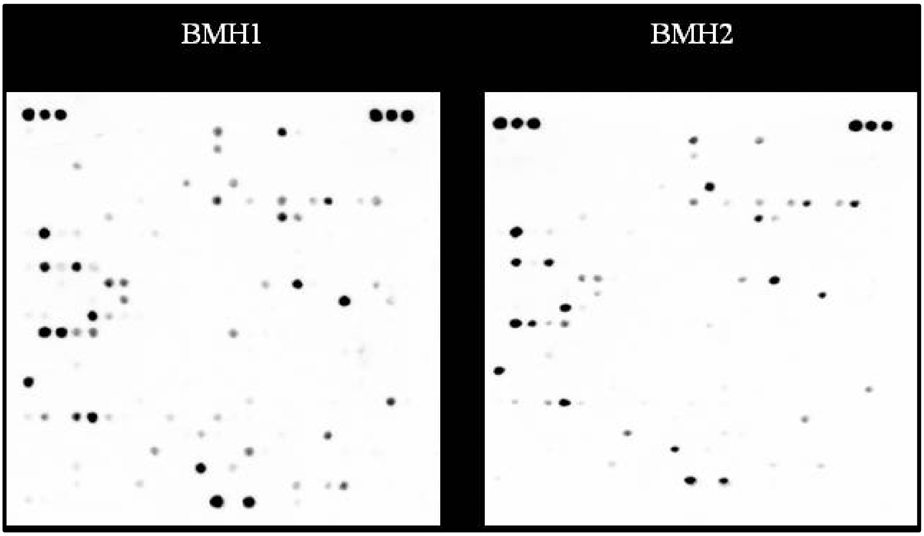
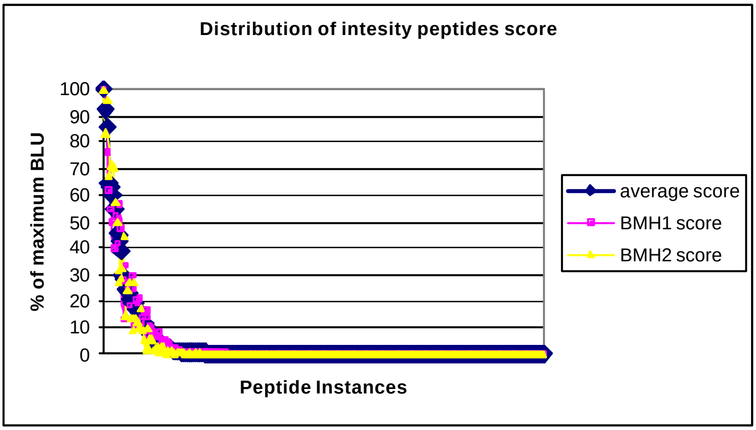


Figure 6 : Yeast phosphoproteome scanning against BMH1 and BMH2 probe.

Also in this experiment, if we compare the quantitative results of BMH1 and BMH2, we observe an almost identical pattern; we establish an arbitrary threshold of BLU units to separate good ligand from non-ligand. In order to normalize the results of the two membranes (the BMH1 membrane has visibly a stronger signal than the BMH2 one) we set the cut-off value as the 10% of the maximum BLU (see Graph 2), corresponding to BLU values of 22.000 and 17.000 respectively in BMH1 and BMH2 phosphoproteome scanning respectively. Out of 287 peptides known to be phosphorylated *in vivo*, 26 had a score higher than the threshold and were therefore classified as putative ligand (see below Table 4).



Graph 2 : Analysis of phosphospetides BLUs when screened against the 14-3-3 domain of BMH1 and BHM2. The two 14-3-3 domains show identical specificity, thus we set a cut off value corresponding to 10% of each set maximum BLU value.

We next tested our prediction tool against the same set of 287 peptides by first filtering with the regular expression, and then assigning a score to each peptide by summing the values associated to each residue at each position of the 14-3-3 position specific scoring matrix (the PSSM score). The prediction results are displayed in two tables: in the first (see Table 4) we display the 26 positive peptides experimentally proven to be high affinity ligands. In the second table (see Table 5) we show the top 26 peptides according to their PSSM score. 206 out of the 287 peptides do not match the sequence requirements summarized in the regular expression. As reported in table 4, only one of the positive peptides is filtered out erroneously, whereas potential false positives are appropriately excluded from the prediction (see also Table 5). As it is also evident in Table 4, the sequences of the positive peptides have no striking regularities that could be described by a simple regular expression representing the binding motif. In most cases, the 14-3-3 PSSM succeeded in associating high scores to the experimentally proven ligand peptides.

Prediction rank	**PSSM score**	**Regular expression**	Peptide	ORF	Average BLU	% BLU max	Experimental rank
24	**0.509**	**MATCH**	RMAHRSsLSSLSN	YCR023C	152919	100.00	1

11	**0.534**	**MATCH**	ARFSRRsDSGVHS	reg1	141543.5	92.56	2
2	**0.662**	**MATCH**	RKRRASsLKA---	rps6a	131403	85.93	3
5	**0.577**	**MATCH**	RRLSSLsEFNDPF	nth1	98596.5	64.48	4
3	**0.634**	**MATCH**	PKFRRAsLNSKTI	yak1	96201	62.91	5
13	**0.530**	**MATCH**	KARRSMsLLGYRA	myo3	91887	60.09	6
7	**0.551**	**MATCH**	ERRSMVsSPNRYV	spc98	84130.5	55.02	7
91	***0.367***	***MATCH***	*RTSSSMsVGNNKK*	*YMR261C*	*84062*	*54.97*	*8*
20	**0.502**	**MATCH**	SQSRSRsSVMFKS	sra1	69896.5	45.71	9
19	**0.503**	**MATCH**	HRSSLSsLSNQRC	YCR023C	68319.5	44.68	10
14	**0.527**	**MATCH**	LSSKSHsVPALNT	YNL321W	58950	38.55	11
60	***0.414***	***MATCH***	*ALKVRTsATFRLP*	*rpl25*	*46136.5*	*30.17*	*12*
-	***-***	***-***	*SKVYARsVYDSRG*	*eno1*	*43722.5*	*28.59*	*13*
16	**0.520**	**MATCH**	SPLRARsATPTLQ	blm3	37490.5	24.52	14
8	**0.538**	**MATCH**	KVARPLsVPGSPR	gsy2	36775.5	24.05	15
82	***0.378***	***MATCH***	*RQKSTSsYSSGGR*	*akl1*	*35210*	*23.03*	*16*
12	**0.531**	**MATCH**	GTFRRRsSVFENI	syg1	31427	20.55	17
9	**0.536**	**MATCH**	RQRRLSsLSAFND	nth1	30260.5	19.79	18
4	**0.581**	**MATCH**	GRKRSSsSVSLKA	gpd1	29624	19.37	19
17	**0.510**	**MATCH**	SRSRSHsFYKGGH	YGL181W	28791.5	18.83	20
10	**0.534**	**MATCH**	TFRRRSsVFANIS	syg1	25922.5	16.95	21
24	**0.472**	**MATCH**	GHSRASsFARTLA	YML072C	24157.5	15.80	22
26	**0.451**	**MATCH**	SMGRTAsALSRTR	YGR138C	24084.5	15.75	23
25	**0.458**	**MATCH**	QRRRsSYAF	YCR077C	22585.5	14.77	24
16	**0.522**	**MATCH**	KIARPLsVPGSPK	gsy1	17672	11.56	25
62	***0.411***	***MATCH***	*RTATPQsLQGSNK*	*vps13*	*17524.5*	*11.46*	*26*

Table 4: Prediction results sorted according to the experimental rank. We report the 26 highest affinity peptides that have BLUs above the cut off threshold. In the first three columns we report in bold the prediction for each of these positive peptides. The peptides underlined in italics are the 4 false negatives wrongly predicted as low affinity ligands, and have a prediction rank below the 26th position.

PSSM rank	PSSM score	Regular expression	Peptide	ORF	Average BLU	% BLU max	Experimental rank
1	*0.717*	*MATCH*	*YRRRKSsLVVPPA*	*yak1*	*1927.5*	*1.26*	*45*
2	**0.662**	**MATCH**	**RKRRASsLKA---**	**rps6a**	**131403**	**85.93**	**3**

3	**0.634**	**MATCH**	**PKFRRAsLNSKTI**	**yak1**	**96201**	**62.91**	**5**
	0.609		VRKMSFsGYSPKP	YOR175C	5886.5	3.85	35
4	**0.581**	**MATCH**	**GRKRSSsSVSLKA**	**gpd1**	**29624**	**19.37**	**19**
	0.580		PRKRAAsIRARVK	rpl3	12310.5	8.05	29
5	**0.577**	**MATCH**	**RRLSSLsEFNDPF**	**nth1**	**98596.5**	**64.48**	**4**
6	*0.574*	*MATCH*	*IRKRRAsSLKA--*	***rps6a***	*1364*	*0.89*	*49*
	0.555		RKNRSPsPPPVYD	YLR116W	59.5	0.04	83
7	**0.551**	**MATCH**	**ERRSMVsSPNRYV**	**spc98**	**84130.5**	**55.02**	**7**
8	**0.538**	**MATCH**	**KVARPLsVPGSPR**	**gsy2**	**36775.5**	**24.05**	**15**
9	**0.536**	**MATCH**	**RQRRLSsLSAFND**	**nth1**	**30260.5**	**19.79**	**18**
10	**0.534**	**MATCH**	**TFRRRSsVFANIS**	**syg1**	**25922.5**	**16.95**	**21**
11	**0.534**	**MATCH**	**ARFSRRsDSGVHS**	**reg1**	**141543.5**	**92.56**	**2**
12	**0.531**	**MATCH**	**GTFRRRsSVFENI**	**syg1**	**31427**	**20.55**	**17**
13	**0.530**	**MATCH**	**KARRSMsLLGYRA**	**myo3**	**91887**	**60.09**	**6**
14	**0.527**	**MATCH**	**LSSKSHsVPALNT**	**YNL321W**	**58950**	**38.55**	**11**
	0.525		PIRRSDsAVSIVH	YOL059W	21	0.01	97
15	**0.522**	**MATCH**	**KIARPLsVPGSPK**	**gsy1**	**17672**	**11.56**	**25**
16	**0.520**	**MATCH**	**SPLRARsATPTLQ**	**blm3**	**37490.5**	**24.52**	**14**
17	**0.510**	**MATCH**	**SRSRSHsFYKGGH**	**YGL181W**	**28791.5**	**18.83**	**20**
18	**0.509**	**MATCH**	**RMAHRSsLSSLSN**	**YCR023C**	**152919**	**100.00**	**1**
19	**0.503**	**MATCH**	**HRSSLSsLSNQRC**	**YCR023C**	**68319.5**	**44.68**	**10**
20	**0.502**	**MATCH**	**SQSRSRsSVMFKS**	**sra1**	**69896.5**	**45.71**	**9**
21	*0.499*	*MATCH*	*AYRRRKsSLVVPP*	*yak1*	*1712.5*	*1.12*	*47*
22	*0.481*	*MATCH*	*VMKRSAsYTGAKV*	*YDR074W*	*12801*	*8.37*	*28*
23	*0.477*	*MATCH*	*KKSTPVsTPSKEK*	*YHR052W*	*3548*	*2.32*	*41*
24	**0.472**	**MATCH**	**GHSRASsFARTLA**	**YMl072C**	**24157.5**	**15.80**	**22**
25	**0.458**	**MATCH**	**QRRRsSYAF**	**pat1**	**22585.5**	**14.77**	**24**
26	**0.451**	**MATCH**	**SMGRTAsALSRTR**	**YGR138C**	**24084.5**	**15.75**	**23**

Table 5 : Prediction results sorted according to the 14-3-3 PSSM. True positives are in bold; the only 5 false positives are underlined in italics. Only 4 peptides that have high PSSM score and do not match the regular expression are shown in the table, but these are in total 15 peptides that would rank within the first 26 positions if they were not filtered out by the regular expression.

The proposed prediction tool recovers 21 true positives out of the 26 experimentally proven ligand peptides. If we compare the experimental and prediction rank with a cut off on the 26th position, we see that 4 predicted ligand peptides are false negatives (see in Table 4 the experimental positives peptides that are predicted to have low affinity) and 5 are false positives (see in Table 5 the experimental low affinity peptides that are predicted to be good ligand).

Comparison of our predictions with SCANSITE and ELM

In order to compare the performance of our predictions with SCANSITE[1] and ELM[2] predictions, we carried out identical blind test on the same dataset with these two specific tools.

SCANSITE (Obenauer *et al.*, 2003) is based on PSSM derived from experimental data, determined either by phage display or using a degenerated peptide library. Phage display (Scott & Smith., 1990) is a method based on cloning complex peptide libraries in phage vectors expressing the exogenous sequence on their capsides. This phage population then is incubated with the probe domain fixed on a solid support, and the clones expressing ligand peptides are retained. Sequencing the genome of the clones that are selected after this panning procedure leads to the identification of the binding peptide sequences. Similarly, in peptide library experiments (Yaffe & Cantley 2000), degenerate peptides with a single fixed central residue are incubated with a given probe domain. Domain binding peptides are isolated and sequenced as a mixture by Edman degradation. When peptides are sequenced in this manner, one obtains the relative enrichment of each amino acid occurring at each position. Both methods provide sequence information that is normalized to produce a scoring matrix which quantitatively indicates the preference of specific amino acids at each position. Matrices for each analyzed domain can be queried on the SCANSITE public site to scan entire proteomes or a single protein sequence and the server returns high ranking motifs matches. The predictions have three possible levels of stringency (“high”, ”medium”, “low”), based on a comparative scoring approach. A given match is considered high stringency when its score falls within the top 0.2% of all scores calculated when the given motif matrix is applied to a reference set of protein, consisting of all vertebrate proteins in the UniProt database (Bairoch *et al.*, 2005) . The medium and low stringencies are similarly chosen at 1% and 5%

[1] http://scansite.mit.edu

[2] http://elm.eu.org

respectively. Moreover, the user can also customize the protein reference set during the query and, for instance, choose one specie s proteome.

The Eukaryotic Linear Motif server, or ELM server (Punterwoll *et al.*, 2003), is a database of short linear motifs used in eukaryotic organisms for cell compartment targeting, protein-protein interaction, regulation by phosphorylation, acetylation and other post-translational modifications. The collection of motifs characterizing these short functional sites is done by manual curation of the literature, meaning that it is based on experimental data as well. ELMs are represented as regular expressions and associated to extra annotation relative to the motif, such as Gene Ontology (GO)(Gene Ontology consortium, 2001) terms describing cellular compartment, biological process and molecular function, the specific taxa in which it is found and also instances of proteins having true positive matches (sequences having the motif and also its associated function). ELMs are in general simple motifs with a high number of matches across protein sequences, many of which are false positives (sequences having the motif but not its associated function); this is because having the required motif is not a guarantee that a sequence will be a functional site in the biological context. For this reason the aim of the EML server is not only to provide a repository of annotated motifs, but also to provide tools to refine predictions of new instances of an ELM in a query protein. The so called ELM filters are the taxonomy filter, the cell compartment filter and the globular domain filter. The taxonomy filter restricts the space of prediction to the allowed organisms, according to the annotation associated to the ELM. Similarly, the cell compartment filter will filter out the ELM matches belonging to proteins with cellular compartments different from the one associated to the specific ELM. Finally, the globular domain filter based on the GlobPlot (Linding *et al.*, 2003) tool, aims at identifying the matches falling in unfolded regions (inter domain or loop segments), that are more likely to be accessible to functional interactions compared to those belonging to structured domains.

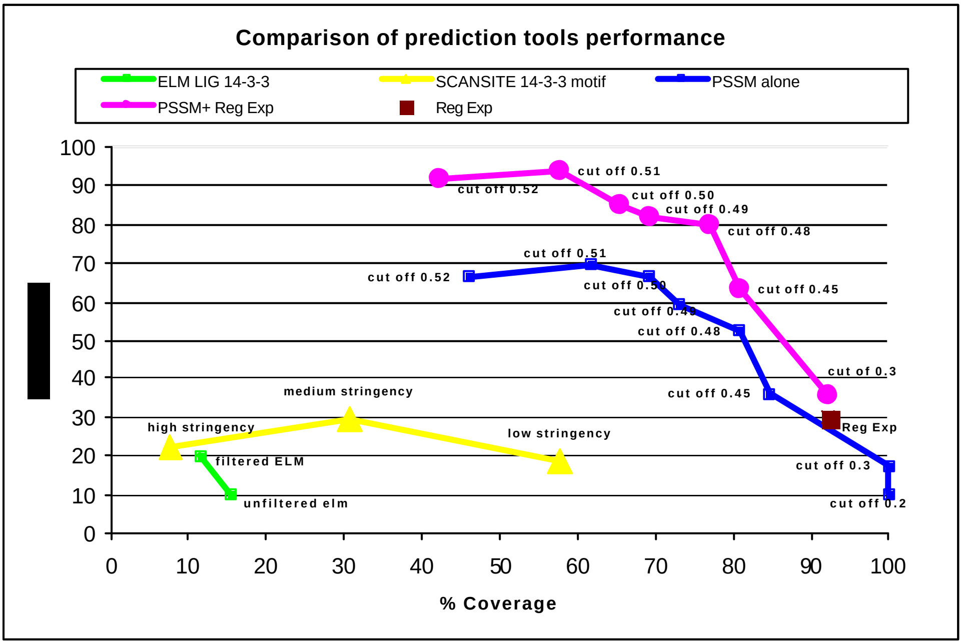
Because both SCANSITE and ELM do not accept peptides as an input sequence, we proceed in the following manner. We submit to SCANSITE an artificial protein sequence, generated by merging all the phosphopeptides of the experimental set, and perform the query in *S. cerevisae* context with all the 14-3-3 motifs simultaneously. We repeat the query at the three levels of stringency and post-process the results, in order to consider only peptides centred on appropriate central phosphoserine.

For ELM we download the “LIG 14-3-3” entry and mimic the server filters locally. The LIG 14-3-3 ELM is annotated as a *S. cerevisae* motif represented by [RHK][STALV].[ST].[PESRDIF] regular expression and observed to be biologically active in nucleus, mitochondrion, cytoplasm and plasma membrane contexts. Subsequently, we download from SGD (Balakrishnan *et al.*, 2005) GO annotation for all yeast ORF and also ask for the GlobPlot prediction for the full yeast proteome (kindly provided by Rune Linding). We then have all elements to grossly imitate the ELM query by running a simple script that, taking the phosphopeptides input sequences, will select those matching the LIG 14-3-3 regular expression, then filter those belonging to ORF that according to GO are located in the ELM subcellular compartment and finally check whether or not the peptide is predicted to be in an unstructured region within its protein full sequence.

In parallel we run our prediction tool *with* and *without* the regular expression filter, also testing different cut-off thresholds of the score calculated from the mutagenesis derived matrix.

All the prediction tools receive as input the 287 phosphopeptides scanned by PepSpot and for each one of them we estimate their capacity to recover the 26 positives peptides that are experimentally shown to be good ligands. To this aim we calculate the accuracy and coverage using the following formula :

Accuracy = (True Positives/ (True Positives+ False Positives)) x100
Coverage = (True Positives/ Experimental Positives) x100



Graph 3 : Accuracy versus coverage plot of the different (SCANSITE, ELM and ours) prediction tool results when tested in their ability to mimic the phosphopeptides selection by BMH1 and BMH2.

Amazingly, SCANSITE and ELM have a very weak performance (see Graph 3) and a surprising lack of true positive recovery. SCANSITE at low stringency correctly predicts only 15 out of 26 positive peptides, whereas the ELM LIG 14-3-3 regular expression matches only 4 of the peptides from the same set. Clearly, both tools are meant to receive real protein query and have been forced to predict binding of short peptides, but we would have expected that these tools could better recognize those primary sequences as potential ligand of 14-3-3 domains.

The approach chosen to derive a prediction tool from mutagenesis data is shown to be successful. The combination of PSSM and a regular expression filtering peptides with residues incompatible with the binding, seems to be very effective at least on this dataset. If we consider a cut-off threshold of 0.50 in PSSM score, the prediction has an accuracy of 94% and a coverage of 58%. Apparently, the “negative” information available from mutagenesis and not accessible to others, is an efficient way of reducing the number of false positives, with almost no loss of true positives. Of course, the advantage of our prediction tool is that the blind test is performed on data collected with the same experimental method used to set up the tool. Potential interferences of the membrane in the interaction between any peptide and the 14-3-3 domain are equally present in both results of the mutagenesis and of the phosphopeptides scan, this could explain part of the performance discrepancies.

PART 2: APPLICATION OF THE COMPARATIVE GENOMIC FILTER AND OTHER VALIDATION STEPS FOR AN IN SILICO 14-3-3 BINDING PEPTIDE SCREENING

Considering the remarkable performances of our prediction tool we decide to use it to search the full yeast proteome for candidate ligands of 14-3-3 domains. In order to increase the accuracy of this *in silico* screening approach and to reduce the number of biologically meaningless predictions, we set up several filters aimed at identifying which ones among a set of interactions either identified *in vitro* or inferred *in silico* are likely to occur *in vivo* and have a functional relevance. The comparative genomic filter, is based on a new approach to protein-protein interaction validation that we propose. Basically we assume that functional binding motifs have a slower mutation rate and we expect to find them unaltered in orthologous proteins in closely related organisms. Following a detailed description of this novel comparative genomic filter procedure, we further present a number of different filters based not only on peptide features but also on protein properties, such as their location, expression and function. Finally, the inferred set of functional partners obtained by applying these filters to the *in silico* screening of 14-3-3 targets is compared with the PepSpot and coimmunoprecipitation experimental results.

Comparative genomic filter

Scope

The main assumption justifying this filter is that if a particular pair of sequences is found to interact in *S. cerevisae* and they are conserved in other genomes, this is more likely to be a biologically relevant interaction. To verify this hypothesis we collect the proteomes of 15 yeast species including *S. cerevisae* and identify 5652 clusters of orthologous proteins. In order to assess whether our assumption is correct, we analyse the alignments obtained from each cluster and assess the level of conservation of PROSITE patterns and ELM motifs. Finally, focusing on 14-3-3 domains we establish a procedure to score potential binding peptides according to their level of conservation.

Building of orthologous protein alignments

The first step for the comparative genomic filter set up is the collection of the ORFs predicted for each yeast genome (see Table 1).

Genome	Genome reference	ORFs sequence source
S. cerevisae	Cherry *et al.,* 1997	http://www.ebi.ac.uk/integr8/
S. paradoxus	Kellis *et al.,* 2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download
S. mikatae	Kellis *et al.,* 2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download
S. bayanus	Kellis *et al.,* 2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download
S. kudriavzevii	Cliften *et al.,* 2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download
S. kluyveri	Cliften *et al.,* 2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download
C. glabrata	Dujon *et al.,* 2004	http://cbi.labri.fr/Genolevures/download.php
D. hansenii	Dujon *et al.,* 2004	http://cbi.labri.fr/Genolevures/download.php
K. lactis	Dujon *et al.,* 2004	http://cbi.labri.fr/Genolevures/download.php
Y. lipolytica	Dujon *et al.,* 2004	http://cbi.labri.fr/Genolevures/download.php
K. waltii	Kellis *et al.,* 2004	http://www.broad.mit.edu/seq/YeastDuplication/
A. gossypii	Dietrich *et al.,* 2004	http://www.ebi.ac.uk/integr8/
C. albicans	Jones *et al.,* 2004	http://www-sequence.stanford.edu/group/candida/
S. pombe	Wood *et al.,* 2002	http://www.ebi.ac.uk/integr8/
N. crassa	Galagan *et al.,* 2003	http://www-genome.wi.mit.edu/annotation/fungi/

Table 1 : 15 yeast species with publicly available genomes considered in our study.

Afterwards we use INPARANOID software (Remm *et al.,* 2001) to automatically determine clusters of orthologues sequences. In the early seventies (Fitch, 1970) homologous sequences, i.e. sequences that have a common ancestor, were classified as orthologous or paralogous according to the following definitions. Orthologues are the homologue sequences belonging to different species that have directly evolved from a single ancestor gene and are believed to share the same function. Whereas paralogues are homologue sequences within the same genome, which also evolved from a common ancestor, but through gene duplication events that may have lead to functional divergence. These evolutionary hypothesis are confirmed either by building phylogenetic trees that is very demanding in terms of computer resources when applied to full genomes or by simple all-versus-all sequence comparison.

The INPARANOID software adopts the latter approach; it requires in input the protein sequence files of 2 species (A and B) and uses BLAST (Basic Local Alignment Tool) to calculate all possible pairwise similarity scores of all proteins within and across proteomes (4 all-versus-all BLAST runs are computed A-A, B-B-, A-B and B-A). The scores that are above the threshold of similarity (50 bits, where the bit is a normalization of the E-score into logarithmic scale) with identical residues distributed over more than 50% of the sequence pair length, are used to cluster the sequences. Each cluster has 2 main orthologues corresponding to best reciprocal hits across the 2 proteomes. Moreover, INPARANOID authors distinguish two classes of paralogues: out-paralogues, whose duplication happened before the speciation event and in-paralogues, duplicated after the speciation event and thus genuine co-orthologues. Thus for each orthologue group in-paralogues for both species are clustered within specific similarity limits (see figure1) defined as the similarity score between the main orthologues. Each in-paralogue has a confidence value ranging from 0 to 100% that is a measure of similarity of a given sequence with its main orthologue.

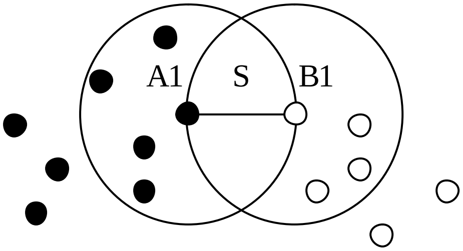


Figure 1 : Clustering of in-paralogues. Each circle represents a sequence from specie A (in black) or specie B (in white). A1 and B1 are the main orthologues determined by reciprocal BLAST best hits, with a similarity score shown as S. In such a similarity space, in-paralogues are closer to their same specie main orthologue than any sequence from the other specie. According to this definition, in-paralogues are inside the circle of diameter S, whereas out-paralogues are outside.

We reasoned that including in-paralogues with 98% of similarity score in our studies will ensure a better coverage of the functional sites conservation across species. In fact, there could be cases where few amino acid substitutions lead to short functional motifs loss in the main orthologue, but not in some of its recent gene replica.
Moreover, INPARANOID software offers the practical advantage of taking as input the two proteomes files under study and by managing automatically the BLAST runs and their parsing, it returns directly a table of orthology groups. We run INPARANOID fourteen times comparing the proteome of *S. cerevisae* with that of the remaining specie. Finally we merged the 14 resulting tables by clustering sequences according to their shared *S. cerevisae* main orthologue.

In this manner we obtain 5652 groups of *S. cerevisae* orthologues (see Table 2). We have not explored whether or not within each group all sequences are orthologues to each other (running INPARANOID with all pairwise combinations of proteomes), which is beyond our scope, we just collect sequences that are highly homologous to the *S. cerevisae* ORFs and likely to have same functions.

Genome	**N° of ORFs**	**N° main orthologues**	**N° in-paralogues**	**% of ORFs being main orthologues**
S. cerevisiae	6222	5652	1806	90.84
S. paradoxus	8955	5415	204	60.47
S. mikatae	9057	5189	134	57.29
S. kudriavzevii	3768	3530	109	93.68
S. bayanus	9424	5149	114	54.64
C. glabrata	5272	4343	118	82.38
K. lactis	5331	4298	75	80.62
S. kluyveri	2968	2753	59	92.76
K. waltii	5230	4271	80	81.66
A. gossypii	4713	4125	48	87.52
D. hansenii	6896	1009	101	14.63
C. albicans	9256	3571	1759	38.58
Y. lipolytica	6666	2483	310	37.25
S. pombe	4931	2780	292	56.38
N. crassa	10082	2861	78	28.38

Table 2 : Contribution of the 15 yeasts proteomes to the *S. cerevisae* centred orthology group. Species are sorted according to their divergence from *S. cerevisae* (see Figure 2).

The number of main orthologues retained for each specie roughly parallels to the rDNA phylogeny (Souciet *et al.*, 2000) reported in figure 2 and is consistent with other comparative studies among the same yeasts (Gaillardin *et al.*, 2000). Gaillardin and co-workers found that 600 genes are conserved in several species but missing in *S. cerevisae*, underlying that a part of the yeast diversity results from species specific gene loss or duplication. Moreover, they analyse sequence conservation in the different gene functional classes, showing that carbohydrate transport and metabolism genes are the most variable classes. Thus, the adaptation of yeast to different environments is not only due to gene loss or duplication but also to rapid evolution of genes and regulatory factors dedicated to sugar metabolism. They also indicate that *Debaryomyces hansenii* and *Yarrowia lipolytica* have a more divergent gene sets than expected from the rDNA phylogenetic analysis. This is consistent with the extreme adaptations of these species, *D. hansenii* being a marine osmostolerant yeast and *Y. lipolytica* a strictly respiratory saprophytic organism.

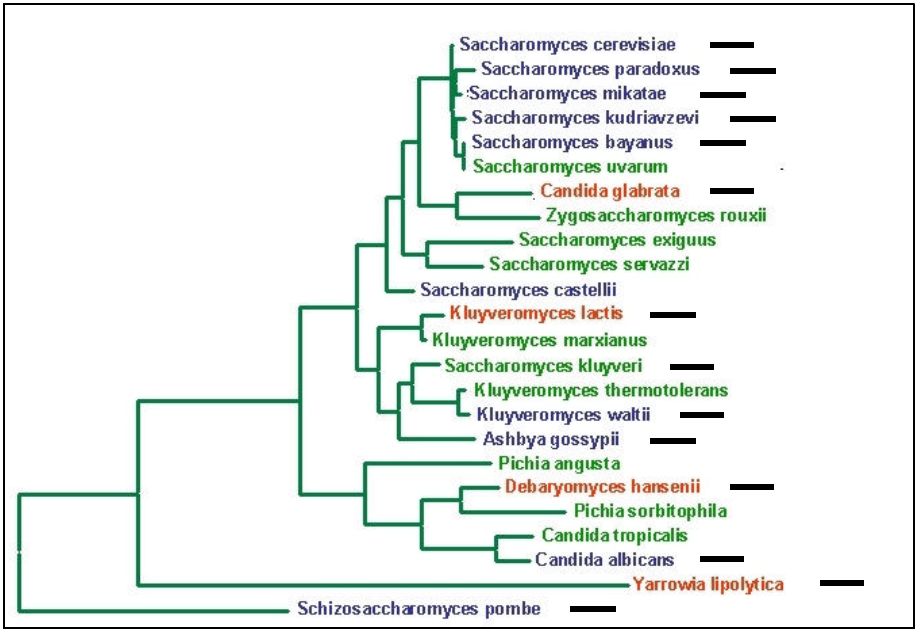
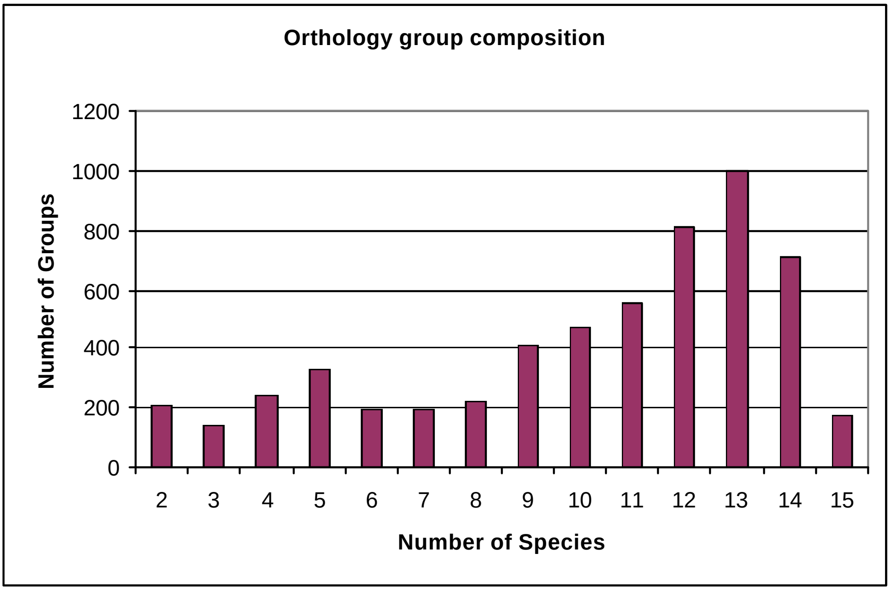


Figure 2 : Cladogram constructed from the coding sequences of the 25S rRNAs using the maximum parsimony built by Souciet and co-workers (Souciet *et al.*, 2000). We add black arrows to the species we have taken into consideration in our studies. *Neurospora crassa* is not included in the tree but, as it belongs to the *Pezizomycotina* class of yeast phylum, it could potentially be displayed as an out-group below *S. pombe*.

Although more divergent species contributed poorly to the *S. cerevisae* orthologues clusters, 73% of the groups contain sequences from more than 7 species (see Graph 1).



Graph 1 : Composition of the 5652 *S. cerevisae* centric, orthologues groups in terms of other species contributions.

Finally, in order to use this data to analyse the conservation of a functional site, we automatically generate a multiple-alignment of all sequences within each group using EMMA, a ClustalW based software from the EMBOSS package (Rice *et al.*, 2000).

Proof of concept using PROSITE patterns and ELM motifs

In order to assess whether or not our basic assumption about functional site conservation is true, we retrieve from PROSITE and ELM regular expressions representing protein signature related to a biological function. We then count, for each *S. cerevisae* sequence matching a regular expression how many orthologs from different species also have a match for the same regular expression.

PROSITE (Hulo *et al.*, 2004) is a manually curated database of regular expressions and profile matrices representation of protein families. The database download consists of two text files: prosite.doc and prosite.dat. The prosite.doc file contains the manual annotations about protein families, whereas the prosite.dat is a computer readable structured file summarizing all information about every entry. The prosite.dat file we use is the release 18.0 (12-Jul-2003) and contains 1639 entries (see in Figure 3 an example of an entry).

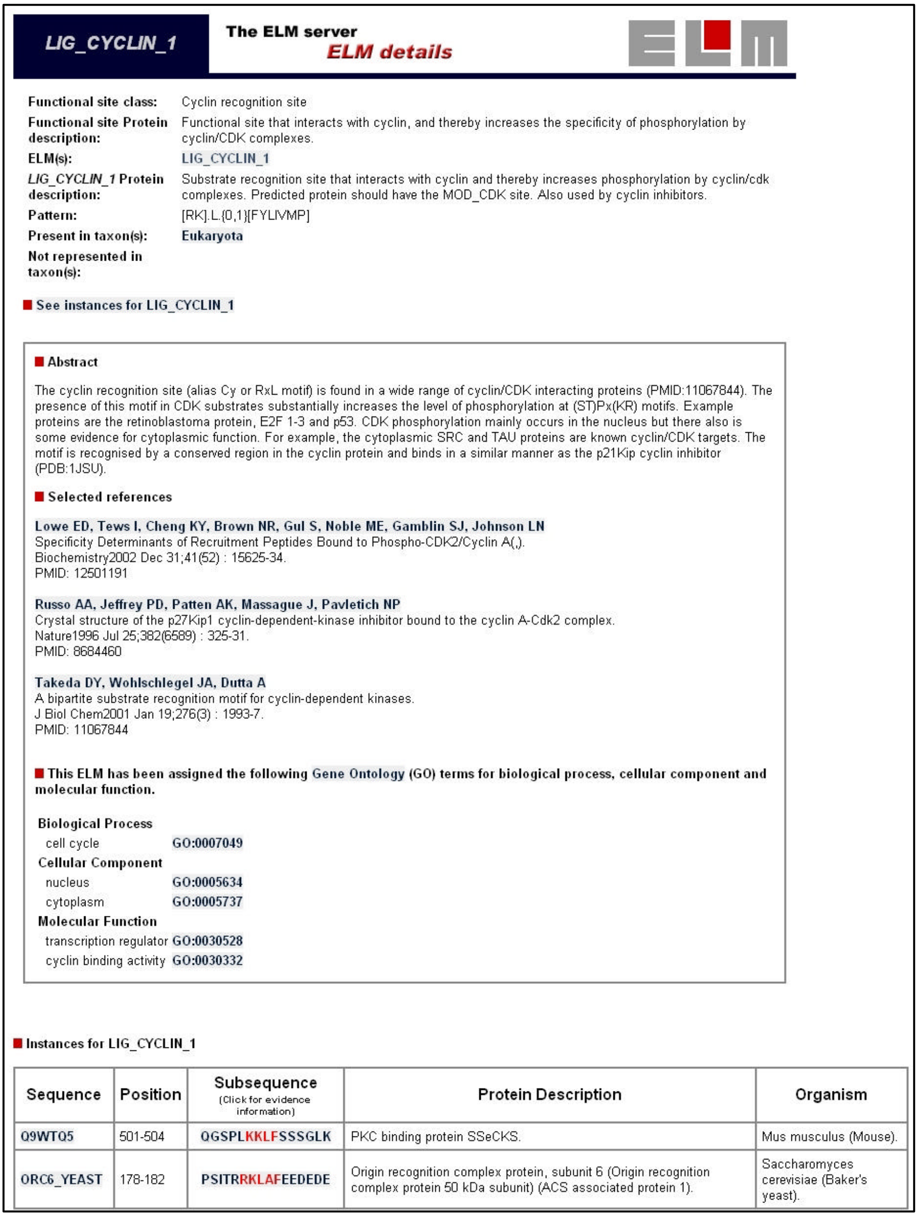
In each entry the first two letters refer to the data contained in the corresponding line. The ID line specifies the name and the type of representation (PATTERN stands for regular expression and MATRIX for profile) of the entry, while the AC line gives its accession number. The biological function common to the protein family is summarized in the DE line.

```
ID WW_DOMAIN_1; PATTERN.
AC PS01159;
DT NOV-1995 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE WW/rsp5/WWP domain signature.
PA W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P.
NR /RELEASE=41.16,130569;
NR /TOTAL=89(62); /POSITIVE=76(49); /UNKNOWN=1(1); /FALSE_POS=12(12);
NR /FALSE_NEG=9; /PARTIAL=1;
CC /TAXO-RANGE=??E??; /MAX-REPEAT=4;
DR O00213, ABB1_HUMAN, T; Q9QXJ1, ABB1_MOUSE, T; P46933, ABB1_RAT , T;
DR Q92870, ABB2_HUMAN, T; O95704, ABB3_HUMAN, T; O35827, ABB3_RAT , T;
DR O95817, BAG3_HUMAN, T; Q9JLV1, BAG3_MOUSE, T; P46942, DB10_NICSY, T;
DR O97592, DMD_CANFA , T; P11533, DMD_CHICK , T; P11532, DMD_HUMAN , T;
DR P11531, DMD_MOUSE , T; P54353, DOD_DROME , T; Q13474, DRP2_HUMAN, T;
DR P22696, ESS1_YEAST, T; O60861, GAS7_HUMAN, T; Q60780, GAS7_MOUSE, T;
DR O55148, GAS7_RAT  , T; P46940, IQG1_HUMAN, T; Q9JKF1, IQG1_MOUSE, T;
DR Q13576, IQG2_HUMAN, T; Q8M9L8, MATK_IMPCA, T; P46934, NED4_HUMAN, T;
DR P46935, NED4_MOUSE, T; Q62940, NED4_RAT  , T; Q13526, PIN1_HUMAN, T;
DR Q9QUR7, PIN1_MOUSE, T; O15428, PINL_HUMAN, T; P33203, PR40_YEAST, T;
DR Q92462, PUB1_SCHPO, T; P39940, RSP5_YEAST, T; O60045, SSP1_NEUCR, T;
DR Q9HCE7, SUF1_HUMAN, T; Q9CUN6, SUF1_MOUSE, T; Q9PUN2, SUF1_XENLA, T;
DR Q9HAU4, SUF2_HUMAN, T; P46939, UTRO_HUMAN, T; Q9H0M0, WWP1_HUMAN, T;
DR Q8BZZ3, WWP1_MOUSE, T; O00308, WWP2_HUMAN, T; Q9DBH0, WWP2_MOUSE, T;
DR Q09685, YA12_SCHPO, T; P46936, YAP1_CHICK, T; P46937, YAP1_HUMAN, T;
DR P46938, YAP1_MOUSE, T; P43582, YFB0_YEAST, T; P46941, YLE5_CAEEL, T;
DR P34600, YO61_CAEEL, T;
DR P11530, DMD_RAT , P;
DR Q9H4Z3, CT67_HUMAN, N; P59114, CT67_MOUSE, N; Q8WYQ5, DGR8_HUMAN, N;
DR O04425, FCA_ARATH , N; Q9C0H5, KG88_HUMAN, N; P59281, KG88_MOUSE, N;
DR O74448, PIN1_SCHPO, N; Q9H4B6, SAV1_HUMAN, N; Q8VEB2, SAV1_MOUSE, N;
DR P40318, SSM4_YEAST, ?;
DR Q9P2S6, AKY1_HUMAN, F; P53868, ALG9_YEAST, F; P12807, AMO_PICAN , F;
DR Q26307, ANA_DROME , F; Q12647, GUNB_NEOPA, F; P47332, LGT_MYCGE , F;
DR P75547, LGT_MYCPN , F; Q9UHC1, MLH3_HUMAN, F; Q00019, RHGB_ASPAC, F;
DR Q07307, UAPA_EMENI, F; P48777, UAPC_EMENI, F; P53076, YGX7_YEAST, F;
3D 1EG3; 1EG4; 1F8A; 1I5H; 1I6C; 1I8G; 1I8H; 1JMQ; 1K9Q; 1K9R; 1O6W; 1PIN;
DO PDOC50020;
```

Figure 3 : The WW_DOMAIN_1 PROSITE entry.

The regular expression is given in the PA line while MA lines refer to profiles. NR (numerical results) lines contain statistics describing the ability of the pattern to discriminate family members from other unrelated SWISS-PROT sequences. By manual cross-check of SWISS-PROT functional annotation, PROSITE curators can specify for all the sequences that match the pattern (/TOTAL) the number of, true positives (/POSITIVES), false positives (/FALSE_POS) or uncertain hits (/UNKNOWN). They also collect from SWISS-PROT the cross-references of the full sequences (/FALSE_NEG), or fragments (/PARTIAL) that are missed by the pattern but are known to have the family biological function. The DR (Database Reference) lines list the SWISS-PROT entries that are picked or missed by the pattern. SWISS-PROT accession numbers followed by "T" are the single instances of the true positives, while "N" stands for false negatives, "F" for false positives and "?" for unknown.
From the proside.dat file we automatically collect 59 pattern entries (i.e. only those having a PATTERN, discarding the entries with a MATRIX representation) that have at least one true positive and one false positive matches in *S. cerevisae* SWISS-PROT sequences. The 496 *S. cerevisae* sequence matches, relative to the 59 patterns, consist of 295 true positives and 201 false positives.

As already mentioned at the end of part1 (see "Comparison of our predictions with SCANSITE and ELM"), ELM is a database of short linear motifs associated to a biological function. ELMs are short stretches of contiguous residues, that do not characterize protein families as PROSITE patterns do, but are nevertheless related to a biological activity. In figure 4 we report an ELM entry that stores a short description of the biological function and the regular expression representing the associated motif. Moreover, every entry reports possible taxonomical restrictions, the GO terms related to the motif and in some cases, ELM instances. The instances are the sequences experimentally demonstrated to have the ELM function, and are equivalent to PROSITE true positive matches. ELM database does not collect false positives, the motifs being low complexity sequence signatures, they can have very high number of matches (LIG_CYCLIN_1 for instance has 18.670 matches in *S. cerevisae* proteome only) and very few are true positives. Therefore, if a sequence is not listed as true positive in the ELM instances we consider it as a false positive.



LIG_CYCLIN_1

The ELM server
ELM details

Functional site class:	Cyclin recognition site
Functional site Protein description:	Functional site that interacts with cyclin, and thereby increases the specificity of phosphorylation by cyclin/CDK complexes.
ELM(s):	LIG_CYCLIN_1
***LIG_CYCLIN_1* Protein description:**	Substrate recognition site that interacts with cyclin and thereby increases phosphorylation by cyclin/cdk complexes. Predicted protein should have the MOD_CDK site. Also used by cyclin inhibitors.
Pattern:	[RK].L.{0,1}[FYLIVMP]
Present in taxon(s):	Eukaryota
Not represented in taxon(s):	

■ See instances for LIG_CYCLIN_1

■ **Abstract**

The cyclin recognition site (alias Cy or RxL motif) is found in a wide range of cyclin/CDK interacting proteins (PMID:11067844). The presence of this motif in CDK substrates substantially increases the level of phosphorylation at (ST)Px(KR) motifs. Example proteins are the retinoblastoma protein, E2F 1-3 and p53. CDK phosphorylation mainly occurs in the nucleus but there also is some evidence for cytoplasmic function. For example, the cytoplasmic SRC and TAU proteins are known cyclin/CDK targets. The motif is recognised by a conserved region in the cyclin protein and binds in a similar manner as the p21Kip cyclin inhibitor (PDB:1JSU).

■ **Selected references**

Lowe ED, Tews I, Cheng KY, Brown NR, Gul S, Noble ME, Gamblin SJ, Johnson LN
Specificity Determinants of Recruitment Peptides Bound to Phospho-CDK2/Cyclin A(,).
Biochemistry2002 Dec 31;41(52) : 15625-34.
PMID: 12501191

Russo AA, Jeffrey PD, Patten AK, Massague J, Pavletich NP
Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex.
Nature1996 Jul 25;382(6589) : 325-31.
PMID: 8684460

Takeda DY, Wohlschlegel JA, Dutta A
A bipartite substrate recognition motif for cyclin-dependent kinases.
J Biol Chem2001 Jan 19;276(3) : 1993-7.
PMID: 11067844

■ **This ELM has been assigned the following Gene Ontology (GO) terms for biological process, cellular component and molecular function.**

Biological Process	
cell cycle	GO:0007049
Cellular Component	
nucleus	GO:0005634
cytoplasm	GO:0005737
Molecular Function	
transcription regulator	GO:0030528
cyclin binding activity	GO:0030332

■ **Instances for LIG_CYCLIN_1**

Sequence	Position	Subsequence (Click for evidence information)	Protein Description	Organism
Q9WTQ5	501-504	QGSPLKKLFSSSGLK	PKC binding protein SSeCKS.	Mus musculus (Mouse).
ORC6_YEAST	178-182	PSITRRKLAFEEDEDE	Origin recognition complex protein, subunit 6 (Origin recognition complex protein 50 kDa subunit) (ACS associated protein 1).	Saccharomyces cerevisiae (Baker's yeast).

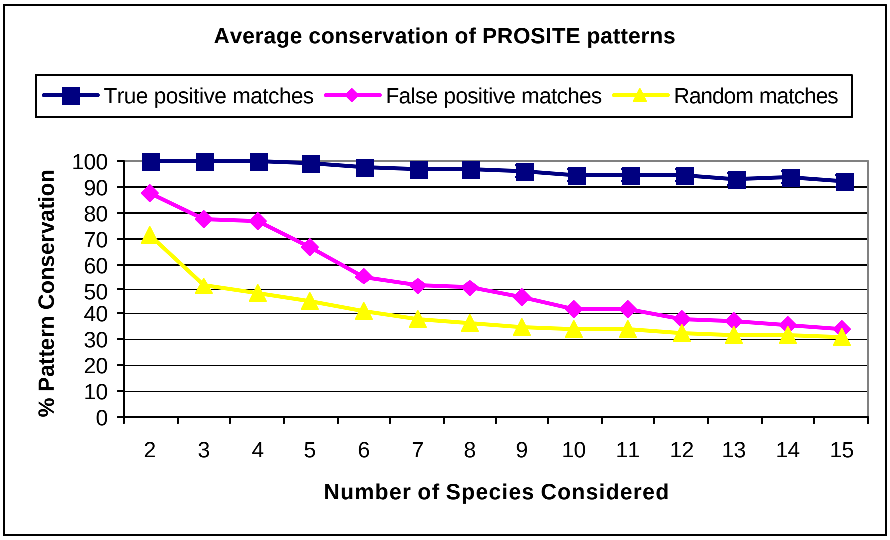
Figure 4 : ELM entry for the LIG_CYCLIN_1 motif.

Unfortunately we could only collect 4 ELMs having ELM instances in *S. cerevisae* proteins: LIG_CYCLIN_1, LIG_AP_GAE_1, LIG_PCNA and LIG_MAD2.

We now have all the inputs necessary to verify whether sequence conservation is a criterion that properly discriminates between true positive and false positive matches of a regular expression associated to a biological function. Moreover to quantify the "noise" conservation due to the evolutionary proximity of the chosen yeasts species, we include in our analysis random, functionally meaningless, patterns. Random patterns are generated inverting the reading direction of the biologically meaningful regular expressions (for instance an AxC motif is inverted to CxA). We measure the conservation of a pattern or motif match in a specific *S. cerevisae* ORF, by automatically counting how many other species still have a match in the ORF orthologues alignment. Because in some alignment the same species can be represented by more than one ortholog (main ortholog and in-paralogs, whenever present) we decide to count the matches per species and not per sequence. Moreover, we impose that the orthologue matches fall in range boundaries dictated by the position of the signature in the *S. cerevisae* sequence. By querying with a regular expression and matching ORF, our procedure returns a percentage of species that keep the pattern unaltered in at least one sequence as a fraction of all the species present in the ORF ortholog alignment. When a pattern or a motif has many matches the "conservation" associated to it is the arithmetic average of all its matches.

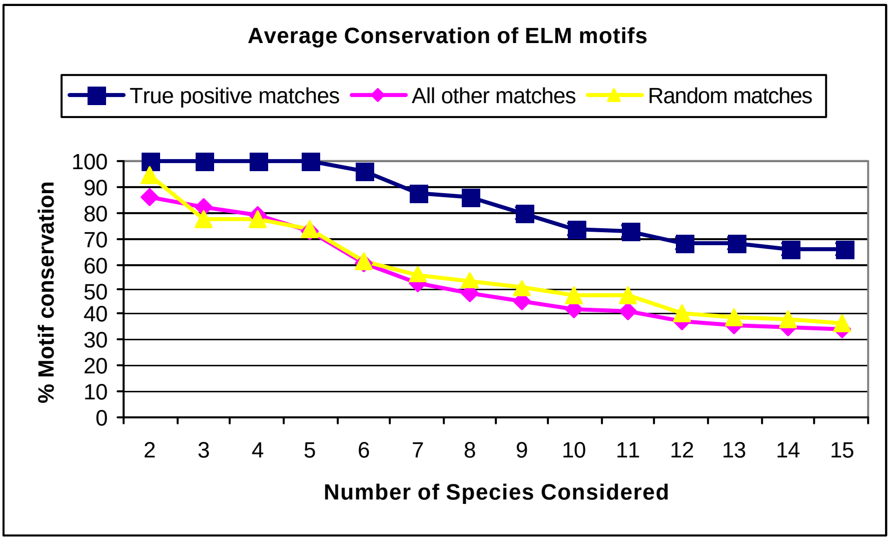
Then for each PROSITE and ELM regular expression (see materials and methods section about protein functional patterns for other details) we analyse independently the conservation of the true positives, or false positives and of all the matches of the randomized pattern. In order to observe the effect of phylogenetic divergence on pattern conservation, we repeat the measures by increasing stepwise the number of species considered in the alignments. The species are added in an order corresponding to their phylogenetic distance from *S. cerevisae*. As shown in graph 2 and graph 3, our analysis confirms that the conservation of peptide sequences in an ortholog alignment decreases more rapidly at increasing phylogenetic distance in random peptides than in peptides matching a functional pattern. Although the difference is less important in the case of the ELM motifs, true positives matches are on average much more conserved than false positives.

One possible explanation to the weaker conservation of ELMs, are possible cases of co-evolution of molecular interfaces underlying a biological function.



Graph 2 : Average conservation of the 59 PROSITE patterns having true positives and false positives matches in *S. cerevisae.*

Graph 3 : Average conservation of the 4 ELMs having true positives instances in *S. cerevisae.*



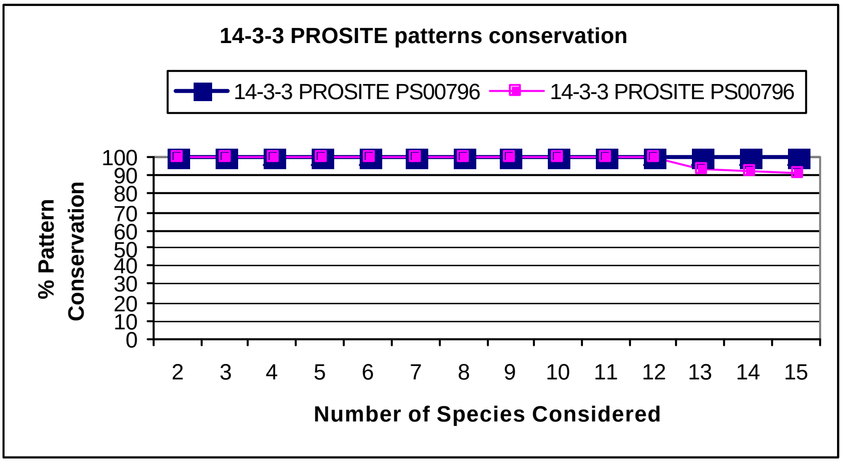
Co-variation of residues or “correlated mutations” events where two interacting partners display at their interface different residue combinations in different species, have been reported in the literature (Pazos & Valencia, 2002). It is reasonable to assume that co-evolution is likely to affect in a stronger manner ELMs in comparison with PROSITE patterns. In fact PROSITE entries mostly define complex long patterns generally characterizing structured protein domains. For this reason a PROSITE pattern does not exclusively report residues directly linked to its specific function, but represents a general sequence signature including residues critical for its structural features. Whereas ELMs, by definition are short linear stretches of sequences experimentally shown to be target of some enzyme or localization signal or just ligand of some other protein. Thus, the residues reported in an ELMs are likely to be all involved in the molecular function much more than the elements of a PROSITE pattern.

Scoring the conservation of putative 14-3-3 ligands

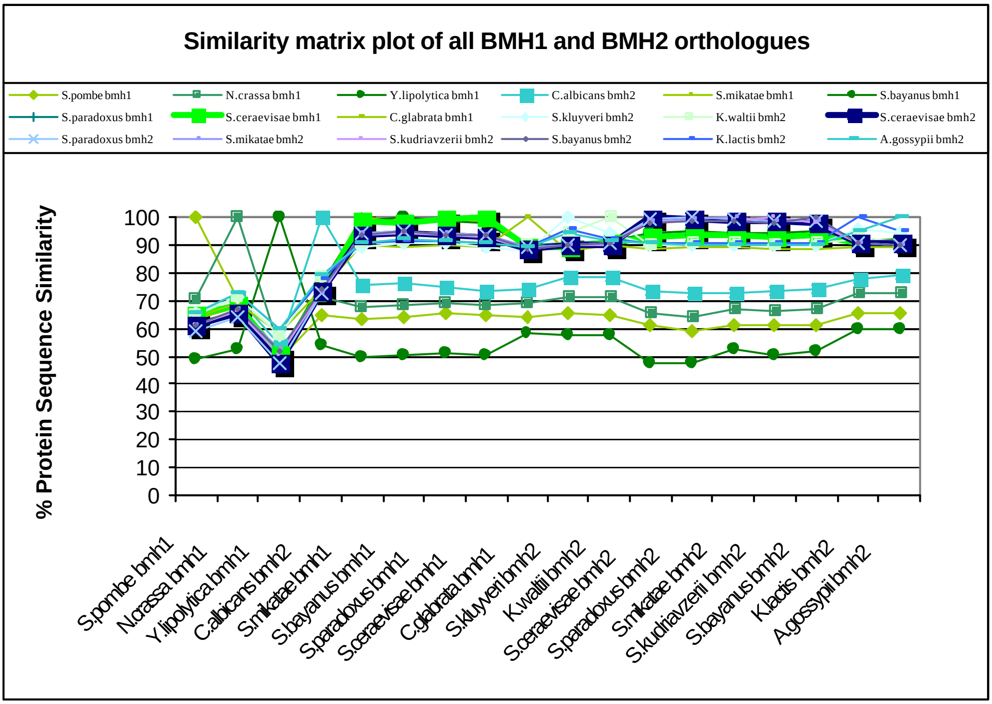
Having acquired evidence that the pattern conservation criterion properly discriminates biologically active sequences, we decide to use it to filter false positive from our prediction of 14-3-3 ligands. However, we have to exclude the eventuality that 14-3-3 domains and their ligands could undergo co-evolution in the range of species we take into consideration. If this happen we might not find the same binding consensus that we have experimentally established in *S. cerevisae.* Although the two patterns representing 14-3-3 domains in the PROSITE database are both almost perfectly conserved in all the orthologs of BMH1 and BMH2 (see Graph 4), we perform a more detailed analysis of the full sequences conservation of the 14-3-3 domains.

We focus on the orthologues alignments generated for BMH1 and BMH2, as expected BMH1 is an in-paralogue of BMH2 and vice-versa. Thus we merge the two sequences files (un-aligned sequences) and align them with all the 14-3-3 proteins of the fifteen yeast species that we have considered. Finally this alignment is submitted to PROT DIST, a software from the PHYLIP package (Felsenstein, 2004) that calculates protein sequence similarity according to the standard amino acid replacement methods (Dayhoff *et al.*, 1972). The all-versus-all comparison results are reported in graph 5. With the exclusion of *N. crassa*, *S. pombe*, *Y. lipolytica* and *C. albicans* all other sequences have very high similarity. Among these we find BMH1 and

BMH2 from *S. cerevisae* for which we have experimental evidence of their identical binding specificity.



Graph 4 : Conservation of 14-3-3 PROSITE patterns in BMH1 and BMH2 alignments.



Graph 5 : All-versus-all sequence comparison of BMH1 and BMH2 orthologues.

Thus, we decide not to consider *N. crassa*, *S. pombe*, *Y. lipolytica* and *C. albicans* orthologue sequences for validation of 14-3-3 ligands, as we fear their 14-3-3 domain may have different specificities. Furthermore, as *D. hansenii* does not have any orthologue sequence to BMH1 or BMH2, we also exclude this specie from the rest of our study.

Procedure for the *in silico* 14-3-3 binding peptide screening

Our aim is to screen all decapeptides from the *S. cerevisae* proteome that have a central serine, use the 14-3-3 prediction tool derived from the PepSpot mutagenesis experiment, to select potential partners and restrict by other criteria the number of predicted interactors. Together with the comparative genomic filter we set up a number of other selection criteria, based on the current protein interaction validation methods, to increase the accuracy of the prediction of yeast 14-3-3 partners on a proteome scale. We now present each filter, beginning with those that we use to select peptides and then continuing with those applied at protein level to restrict the range of our *in silico* peptide screening.

Selection criteria at the peptide level

The first selection undergone by serine centred decapeptides consists in matching the regular expression and having a high score according to the PSSM derived from 14-3-3 binding data. Furthermore, we use the comparative genomic filter to verify that the regular expression pattern representing not tolerated residues is conserved in all orthologous sequences.

Then we try to assess whether or not a given peptide has the sequence requirements to be phosphorylated. We therefore utilize again ELM resource to retrieve all regular expressions representing serine phosphorylation sites. We find 10 target motifs (see Table 3) describing the features of the main eukaryote kinase substrates. We also collect a larger number of eukaryote kinase target consensus from the literature (see Table 4) (Brinkworth *et al.*, 2003). Regular expressions from both sources show that the protein kinase A targets are more likely to be binding partners of 14-3-3 domains. The remaining protein kinases phosphorylate targets containing residues that are incompatible with 14-3-3 binding since they require negatively charged residues or proline in position +1

ELM ID	Regular expression	Motif description
MOD_CDK	...[**ST**]P.[KR]	CDK motif for Ser/Thr phosphorylation
MOD_CK1_1	S..[**ST**]	CK1 motif for Ser/Thr phosphorylation
MOD_CK2_1	[**ST**]..E	CK2 motif for Ser/Thr phosphorylation
MOD_GSK3_1	..[**ST**]...[ST]	GSK3 motif for Ser/Thr phosphorylation
MOD_PK_1	[RK]..**S**[VI]	PK motif for Ser/Thr phosphorylation
MOD_PKA_1	[RK][RK].[**ST**]	PKA motif for Ser/Thr phosphorylation
MOD_PKA_2	R.[**ST**]	PKA motif for Ser/Thr phosphorylation
MOD_PKB_1	R.R..[**ST**].	PKB motif for Ser/Thr phosphorylation
MOD_PLK	[DE].[**ST**][ILFWMVA]	Polo-like-Kinase motif for Ser/Thr phosphorylation
MOD_ProDKin_1	...[**ST**]P..	MAP Kinase motif for Ser/Thr phosphorylation

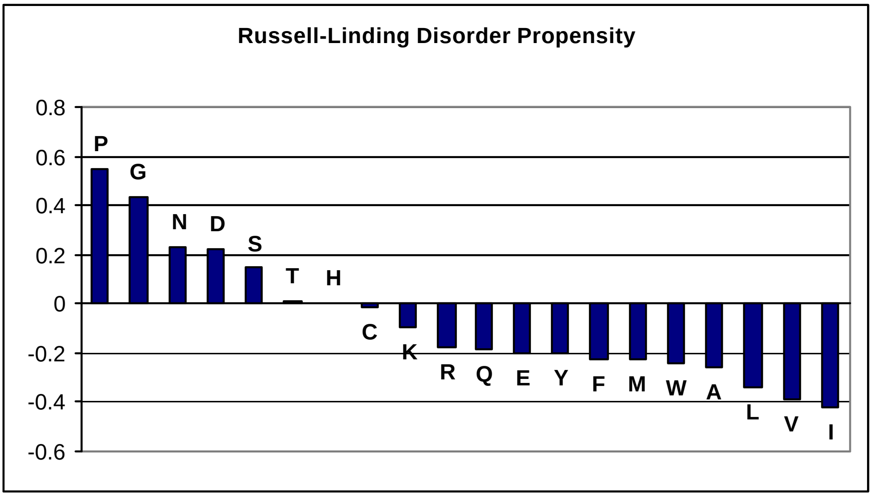
Table 3 : ELM phosphorylation motifs.

Protein Kinase	Consensus derived from known substrates
AKT	R[RTSL][SPRT][**ST**][YSF][PGAST][EAND]
AMPK	R[SNT][MEQN][**ST**][FIGK][LAI][HFLA]
ARK	[NTK][NVT][DS][**ST**][ENDQ][EDNG][RQEDN]
CaMKII	[RKG][QARKLS][AQLG][**ST**][VLIF][SADG][SEMD]
CDK2	[TL][PVLSH][LATS][**ST**]P[PR][KRL]
CDK5	[EG][TVH][KA][**ST**]P[VPE]K
CHK2	R[STI][FHKP][**ST**][DFM][LVPS][WLKE]
CK1	[SDET][SEAV][SLDE][**ST**][ELSVI][ESTD][ESGD]
CK2	[DSE][ESD][EGS][**ST**][DE][ED]E
ERK1	[PAVTL][PV][LT][**ST**]P[PSR][PARKFG]
PHK	R[AQSTL][ILR][**ST**][VITA][RHY][RKFS]
PKA	[RK][RK][SLRGAP][**ST**][LSVR][STPV][SAVGE]
PKC	[RKS][RKAG][LSAR][**ST**][FLVRK][RKAS][RKS]
PKG	R[RK][RLI][**ST**][RASIK][SALK][EPT]
S6K	R[RAS][LS][**ST**][SVL][SRL][SRAG]
SLK1	[LA][AV].[**ST**][FL][TA][TG]

Table 4 : Substrate specificity of some protein kinases retrieved from the literature.

This set of kinase target patterns is not complete, we observed that 20% of the naturally phosphorylated peptides (used in our predictor blind test) do not match any of the regular expression. However, the usage of these regular expressions allow us to screen a full proteome avoiding false positive matches by applying the comparative genomic filter. In fact whatever phosphorylation pattern is matched by a peptide under analysis, we also check its conservation in the orthologous sequences. In this way we first identify sites that show simultaneously binding capacity for 14-3-3 domains and target sequence for Ser/Thr phosphorylation, and then select those that are likely to be true functional sites requiring the high conservation of these two properties in other yeasts species sequences

Moreover we decide to use GlobPlot (Linding *et al.*, 2003) in order to select peptides that belong to "disordered regions". In this way we exclude the sites that belong to some conserved region but are likely to be buried within the folded protein structure. GlobPlot is a tool simply based on each amino acid propensity to be in an unstructured or structured region (see Graph 6).



Graph 6 : Amino acid propensity to be in unstructured protein regions.

These propensity values are calculated by counting how many times a given amino acid is found in regular secondary structure (alfa-helices or beta strands) or outside of them (random coil, turns, or loops), in a non redundant protein structure set (Velankar *et al.*, 2005). When the first derivative of the propensity curve obtained scoring a full length protein is positive in a window of more than 4 residues, this is predicted to be a disorder region, whereas when it is negative for more than 30 amino acid a globular structured region is inferred. GlobPlot successfully predicts ordered region corresponding to SMART (Letunic *et al.*, 2004) domains and disordered stretches in the inter-domain segments (see Figure 5). Moreover GlobPlot finds short regions (7-14 residues) nested within SMART domains.

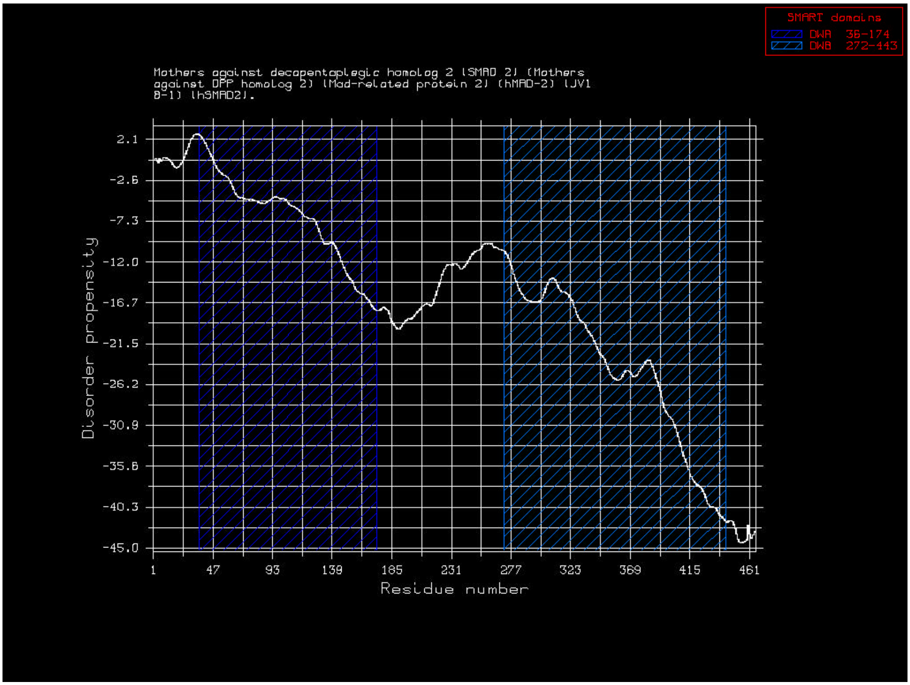


Figure 5 : Disorder propensity plot of the human transcription factor SMA2_HUMAN. GlobPlot downhill regions overlap two SMART structured domains, small peaks within them correspond to the turns between helices and strands.

GlobPlot predictions for single sequences are available online[1] but the full *S. cerevisae* proteome we need for our *in silico* screening has been kindly provided by R. Linding in an upper/lower case code for (upper case residues corresponding to disordered regions). In our studies we have

[1] http://globplot.embl.de/

applied the rule indicated by the author and considered a peptide as potentially unstructured when it has at least 4 amino acid stretch predicted to be disordered.

Selection criteria at the protein level

As presented in the introduction, protein interaction data can be validated by checking that the binding partners satisfy basic co-occurrence rules. Physical protein interaction requires that partners have overlapping expression patterns and are available in the same subcellular location. Moreover, as protein interactions underlie molecular mechanisms supporting biological processes, protein partners with a similar functional annotation are more likely to be binding partners.

Considering the co-expression criterion, ideally we would like to identify all the yeast ORFs with an expression profile that is similar to BMH1 and BMH2 in RNA array hybridisation experiments carried out on different conditional time series (cell cycle, shock treatments, growth conditions ect). Kemmeren and co-workers published a comprehensive list of *S. cerevisae* ORFs pairs that have significantly similar RNA expression profiles in a number of different microarray experiments (Kemmeren *et al.*, 2002). Unfortunately only 4 ORFs show to have some correlation with BMH2 transcription (KSP1, SEF1, RIF2 and ECM13) and only SRP1 result to have an expression profile similar to BMH1.

As an alternative we decide to take advantage of the global analysis of protein expression in *S. cerevisae*, that aim at quantifying protein products during log-phase growth (Ghaemmaghami *et al.*, 2003). In this work the authors successfully introduce within yeast chromosome a TAP (Tandem Affinity Purification) cassette on each gene 5' end, and therefore obtain more than 6100 mutants able to express at their natural level, engineered proteins carrying the same purification flag at their C-terminus. 4.463 proteins were found to be expressed during normal growth and their abundance range from 50 molecules to 10^6 molecules per cell. 1.500 proteins do not have a detectable expression. Nearly 1.000 proteins are known to be required only in specialized conditions (sporulation, or alternative metabolisms) and the remaining 500 are stated to be spurious ORFs generated by automatic translation of the genomic sequence. This statement is supported by the excellent overlap of this set of spurious ORFs with the one reported from comparative analysis of genomic sequences of *S. cerevisae* with three very closely related species (*S. paradoxus*, *S. mikatae* and *S. bayanus*) (Kellis *et al.*, 2003). In our *in silico* screening we decide to take into consideration only the 4.463 proteins that have a detectable expression during log-phase growth and include BMH1 and BMH2. This way we first eliminate spurious proteins from our 6.232 *S. cerevisae* protein collection (see the material and methods section "Building a proteomic table to combine *S. Cerevisae* data"), moreover we "normalize" our screening with those performed experimentally that always occur under normal growth conditions. This set of 4.463 proteins includes around 800 well characterized ORFs whose expression is observed in the localisation experiment below but not detected in the present one. The authors state these proteins that may be expressed at a level below the experimental resolution, or undetected for technical reasons.

Similarly, in order to set up a co-localization filter we use the results of a single large scale protein localisation experiment performed in *S. Cerevisae* (Huh *et al.*, 2003). In this case protein fusions with GFP (Green Fluorescent Protein) are generated engineering 6.029 ORFs, and are co-expressed with 12 protein fusions with RFP (Red Fluorescent Protein) acting as markers of distinct subcellular compartments. Each expressed ORF product is assigned to a compartment by inspecting the yeast cells with a fluorescent microscope. In total 4.156 protein products are located and the specific assignments show strong correlation with previous studies performed with different technologies and formalized in GO format at SGD[1]. We choose to use the data from Huh and co-workers instead of GO annotations for subcellular compartments because of the assignment done for BMH1 and BMH2. According to the authors our two favourite proteins result to be both cytoplasmic and nuclear, whereas according to GO annotations they are only nuclear, that is inconsistent with evidences of BMH2 presence in the cytoplasm (Beck & Hall, 1999). Thus we retrieve the full results[2] and identify all the proteins that have a coherent subcellular location with BMH1 and BMH2. When we exclude the proteins that are isolated by membranes from the nucleus and the cytoplasm 2.882 ORFs result to be co-located with 14-3-3 proteins. When we consider only the proteins that are visualized in

[1] http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.sgd

[2] http://yeastgfp.ucsf.edu/allOrfData.txt

the nucleus or in the cytoplasm 2.377 ORFs results to have the same location as BMH1 and BMH2.

Finally we create a GO based filter able to select proteins that share BMH1 and BMH2 biological process (see materials and methods, the gene ontology section). We retrieve the biological process ontology and the GO assignments for all *S. cerevisae* ORFs[1]. We analyse the GO biological process terms assigned to BMH1 and BMH2 and the position of these terms within the ontology tree (see Figure 6). As expected, BMH1 and BMH2 share all their GO processes that are terms ranging from depth 4 to depth 8, and these terms properly reflect the heterogeneous group of biological functions where our two proteins are known to be involved (from signal transduction to metabolism).

Next, we automatically generate a list of ORFs having at least one common term with BMH1 and BMH2 (*i.e*. ORFs annotated with any of the GO term reported in Figure 6), excluding the ORFs that are annotated with the term 'Biological Process Unknown' (GO:0000004). According to the biological process tree, any pair of ORFs always has at least one GO process in common that is the root term itself 'Biological Process' (GO:0008150). For this reason we also record the depth of the common term shared by any ORFs with BMH1 and BMH2, excluding the trivial pairs whose only common parents term is 'Biological Process' at depth 0 of the tree. 3.598 *S. cerevisae* ORFs result to be sharing with BMH1 and BMH2 at least one parent term, that has depth equal or greater than 1. Only 48 ORFs have a common term at depth 1 whereas nearly 3.300 ORFs share a GO process having a depth ranging from 2 to 4 (see Graph 7). This is when we consider for each ORF only its highest depth term in common with BMH1 and BMH2. In fact if one ORF shares a term at depth 8 it also has other common parent terms at each lower depth level, but in graph 7 we do not report these redundancies.

[1] http://www.geneontology.org/

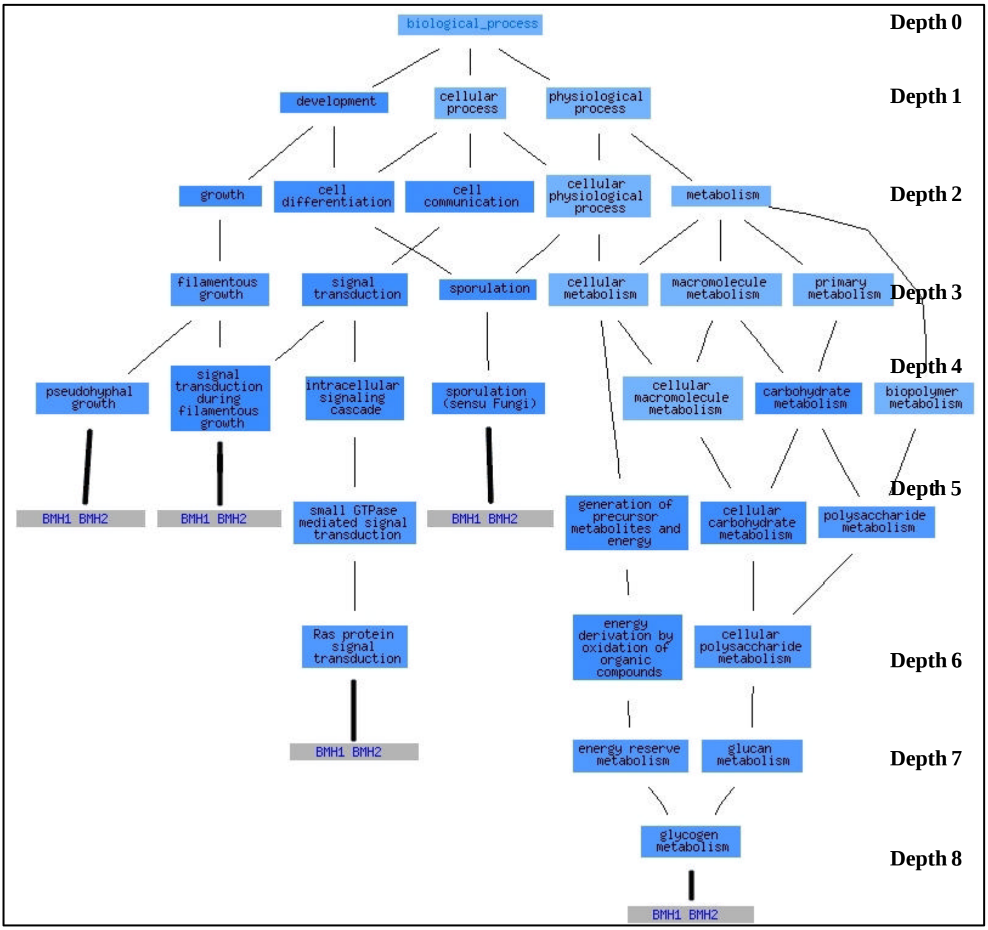
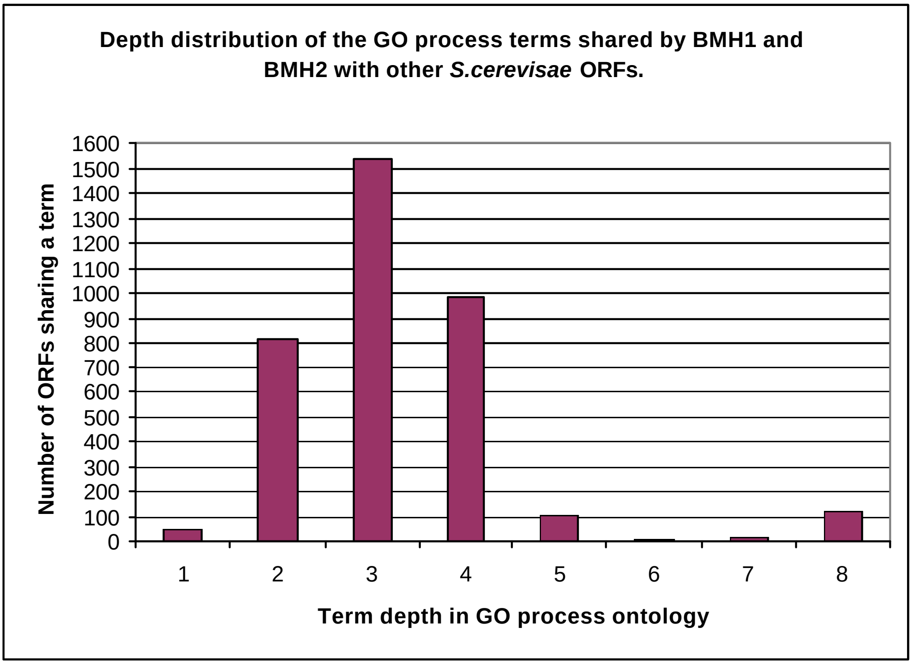

biological_process
development
cellular process
physiological process
growth
cell differentiation
cell communication
cellular physiological process
metabolism
filamentous growth
signal transduction
sporulation
cellular metabolism
macromolecule metabolism
primary metabolism
pseudohyphal growth
signal transduction during filamentous growth
intracellular signaling cascade
sporulation (sensu Fungi)
cellular macromolecule metabolism
carbohydrate metabolism
biopolymer metabolism
BMH1 BMH2
BMH1 BMH2
small GTPase mediated signal transduction
BMH1 BMH2
generation of precursor metabolites and energy
cellular carbohydrate metabolism
polysaccharide metabolism
Ras protein signal transduction
energy derivation by oxidation of organic compounds
cellular polysaccharide metabolism
BMH1 BMH2
energy reserve metabolism
glucan metabolism
glycogen metabolism
BMH1 BMH2
Depth 0
Depth 1
Depth 2
Depth 3
Depth 4
Depth 5
Depth 6
Depth 7
Depth 8

Figure 6 : Part of GO Biological Process tree that includes all the terms assigned to BMH1 and BMH2 (indicated by bold lines). In the right side we indicate the numbering of the tree depth. This image is derived from the graphical output of the GO Term Finder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder) tool, when queried with BMH1 and BMH2.



Graph 7 : Common term depth distribution. The terms that can be shared by BMH1 and BMH2 at each GO process tree level are reported in figure 6.

None of the three protein selection criteria alone can give us indications about specific binding partners of the 14-3-3 domains. Each of them identifies 3 or 4 thousands proteins that may interact with BMH1 or BMH2 according to the selected criteria. However by combining these three filters the number of sequences left to be screened at peptide level drops from the initial 6.232 ORFs to around 2.000 proteins that have all the biological requirements to be 14-3-3 partners. Submitting only these 2000 proteins to the peptide screening (see Figure 7) we ensure that our predicted partners are existing proteins (and not only predicted ORFs), that co-occur in time and location with the 14-3-3 proteins, and also share some common cellular process with them.

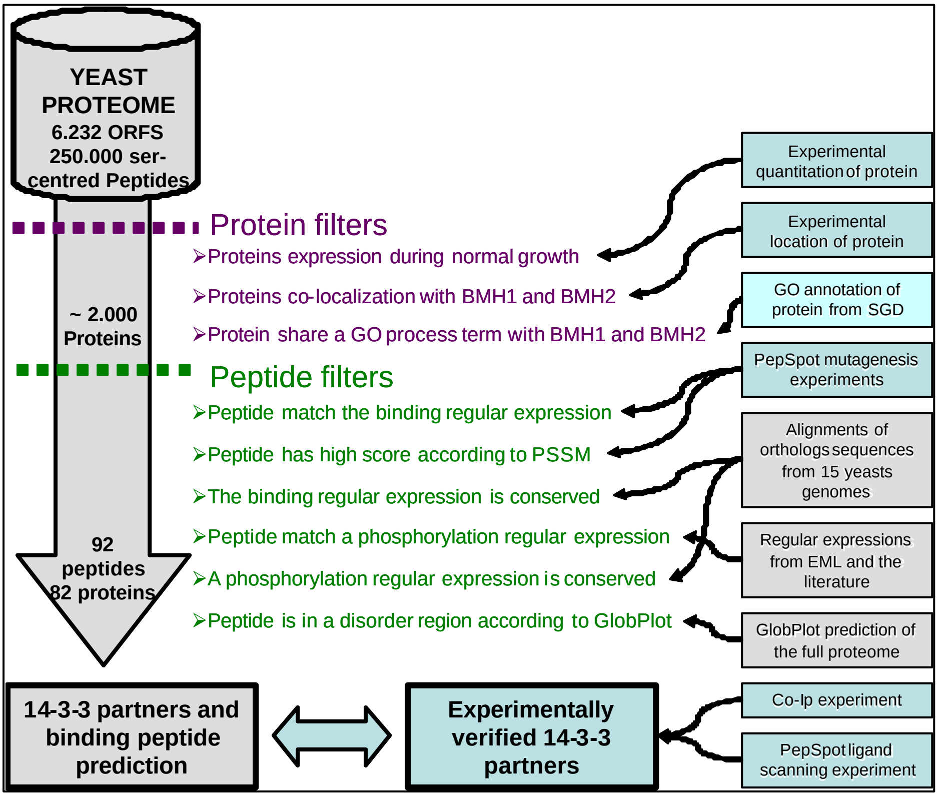

YEAST
PROTEOME
6.232 ORFS
250.000 ser-centred Peptides
~ 2.000
Proteins
92
peptides
82 proteins
Protein filters
➢Proteins expression during normal growth
➢Proteins co-localization with BMH1 and BMH2
➢Protein share a GO process term with BMH1 and BMH2
Peptide filters
➢Peptide match the binding regular expression
➢Peptide has high score according to PSSM
➢The binding regular expression is conserved
➢Peptide match a phosphorylation regular expression
➢A phosphorylation regular expression is conserved
➢Peptide is in a disorder region according to GlobPlot
Experimental quantitation of protein
Experimental location of protein
GO annotation of protein from SGD
PepSpot mutagenesis experiments
Alignments of orthologs sequences from 15 yeasts genomes
Regular expressions from EML and the literature
GlobPlot prediction of the full proteome
Co-Ip experiment
PepSpot ligand scanning experiment
14-3-3 partners and binding peptide prediction
Experimentally verified 14-3-3 partners

Figure 7 : Overall structure of the *in silico* screening of *S. cerevisae* proteome for 14-3-3 mediated interactions. Experimental data are highlighted in blue, whereas predictions and annotations are respectively in grey and green boxes.

Result of the *in silico* screening and comparison with experimental results

In order to assess the performance of our large scale screening, we take advantage of the interaction data concerning BMH1 and BMH2 extracted from the literature and stored in the MINT database. Moreover we carry out in our laboratory further co-immunoprecipitation (CoIp) and PepSpot assays.

We therefore use these experimental interaction results to assess the accuracy of each filter in selecting 14-3-3 domain partners. In this way we can establish appropriate thresholds and refine our final prediction of BMH1 and BMH2 partners.

In vivo detection of *S. cerevisae* 14-3-3 ligands

By querying the MINT database we find 27 proteins that interact with one or both the two yeast 14-3-3 proteins, BMH1 and BMH2 (see Table 5). Because in our studies we cannot observe any divergence in specificity of these two isoforms we consider all their partners as 14-3-3 domain interactors. Most of the interactions have been identified by high throughput screening and only three proteins (KCS1, NTH1 and REG1) have two independent experiments supporting their interaction with 14-3-3 domains.

Nevertheless we repeat a TAP Co-Ip (Tandem Affinity Purification tag Co-Immunoprecipitation) using BMH1 and BMH2 as bait to isolate protein complexes (Panni *et al.*, *in prep*). The tagged 14-3-3 are retained by a calmodulin resin with their partners. The complexes are then disassembled by SDS PAGE (Sodium Dodecyl Sulphate PolyAcrylamide Gel Eelectrophoresis) and the 14-3-3 interacting proteins identified by mass spectrometry. As shown in figure 8, the patterns for BMH1 and BMH2 are very similar with the largest bands being the tagged baits themselves. Moreover these results confirm that BMH1 and BMH2 form homo and heterodimers *in vivo*. In fact BMH1 is fished as interactor when BMH2 is used as bait and *vice versa*. The proteins identified both in the BMH1 and BMH2 CoIp are REG1, RGT2, NTH1, MKS1, HEM15, YPL100C and FAS1, whereas FAS2 is found only in the BMH1, and MYO2 exclusively in the BMH2 experiments. Only 3 of these 14-3-3 partners are already known raising once again the issue of explaining the lack of overlap in independent interaction screening results. However, to test the *in silico* prediction we need information about the peptides responsible for the interaction. Thus we carry out a PepSpot ligand screening, i.e. the analysis of all the serine containing peptides derived from each protein sequence that is known to be a 14-3-3 partner.

Partner	Interaction identification method	Reference
ADR1	flag tag coimmunoprecipitation	Ho *et al.*, 2002
BNR1	flag tag coimmunoprecipitation	Ho *et al.*, 2002
BOI2	flag tag coimmunoprecipitation	Ho *et al.*, 2002
BOP3	two hybrid array	Uetz *et al.*, 2000
CSR2	flag tag coimmunoprecipitation	Ho *et al.*, 2002
CYK3	flag tag coimmunoprecipitation	Ho *et al.*, 2002
ECM13	two hybrid array	Uetz *et al.*, 2000
FUN31	tap tag coimmunoprecipitation	Gavin *et al.*, 2002
GSY2	flag tag coimmunoprecipitation	Ho *et al.*, 2002
KCS1	**two hybrid array**	**Uetz et al., 2000**
KCS1	**flag tag coimmunoprecipitation**	**Ho et al., 2002**
LCB2	tap tag coimmunoprecipitation	Gavin *et al.*, 2002
MSN2	coimmunoprecipitation	Beck *et al.*, 1999
MSN4	coimmunoprecipitation	Beck *et al.*, 1999
NTH1	**tap tag coimmunoprecipitation**	**Gavin et al., 2002**
NTH1	**flag tag coimmunoprecipitation**	**Ho et al., 2002**
NTH2	tap tag coimmunoprecipitation	Gavin *et al.*, 2002
PIK1	tap tag coimmunoprecipitation	Gavin *et al.*, 2002
REG1	**flag tag coimmunoprecipitation**	**Ho et al., 2002**
REG1	**two hybrid/pull down**	**Mayordomo et al., 2003**
RIF2	two hybrid pooling approach	Ito *et al.*, 2001
RTG2	tap tag coimmunoprecipitation	Gavin *et al.*, 2002
SEF1	two hybrid pooling approach	Ito *et al.*, 2001
SNF4	tap tag coimmunoprecipitation	Gavin *et al.*, 2002
SOK1	flag tag coimmunoprecipitation	Ho *et al.*, 2002
SRP1	two hybrid pooling approach	Ito *et al.*, 2001
SVL3	flag tag coimmunoprecipitation	Ho *et al.*, 2002
YAK1	coimmunoprecipitation	Moriya *et al.*, 2001
YFR017C	flag tag coimmunoprecipitation	Ho *et al.*, 2002
YIL028W	flag tag coimmunoprecipitation	Ho *et al.*, 2002

Table 5 : 14-3-3 domain partners stored in the MINT database. Proteins in bold are those whose interaction with BMH1 or BMH2 was observed twice in independent experiments.

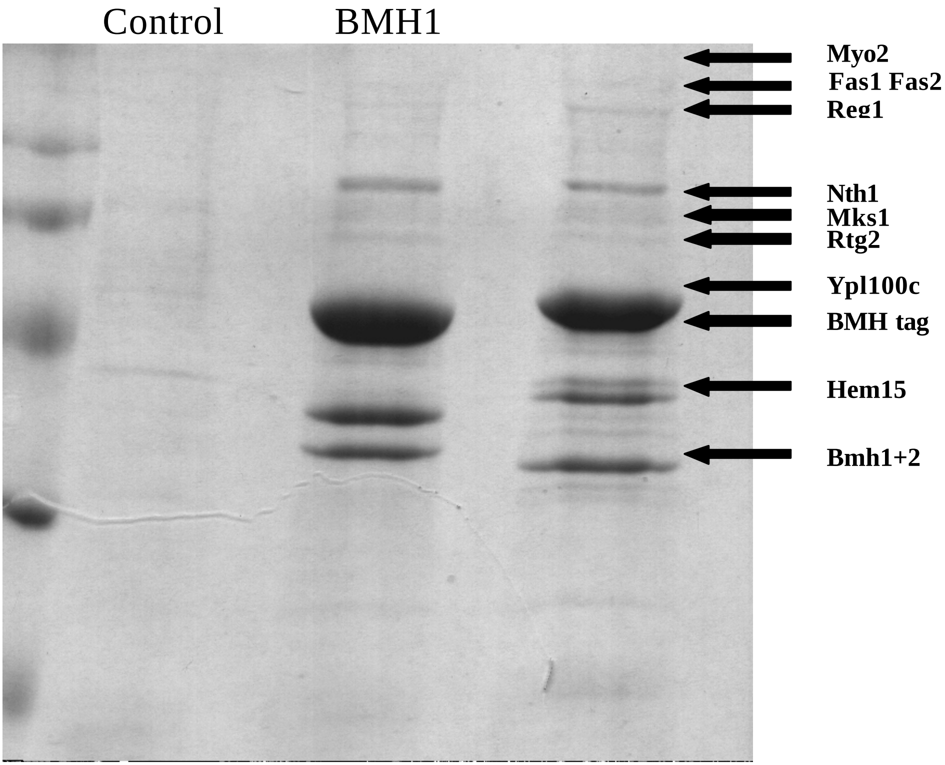


Figure 8 : SDS PAGE of complexes purified using BMH1 and BMH2 as bait.

We include in this screening all the proteins identified in our Co-Ip experiment plus GSY2, MNS2 YAK1 and RTG3, for whom we have both an *in vivo* interaction evidence in MINT and an *in vitro* confirmation in the phosphopeptide PepSpot analysis. From these 13 ligand proteins we derive 1051 peptides having a central serine residue synthesized in their phosphorylated state on membrane and subsequently incubated with BMH1 and BMH2 (see Figure 9). The BLU intensity associated to each of the 1051 peptides show an exponential distribution similar to the one reported previously resulting from the phosphopeptides screening (see Graph2 in part1). Similarly we set a cut off value as the 10% of the maximum BLU score obtain in the screening and identify 69 peptides that we classified as ligands as they have a BLU signal above the threshold. Each of the 13 ligand protein has at least two putative ligand peptides that may mediate complex formation.

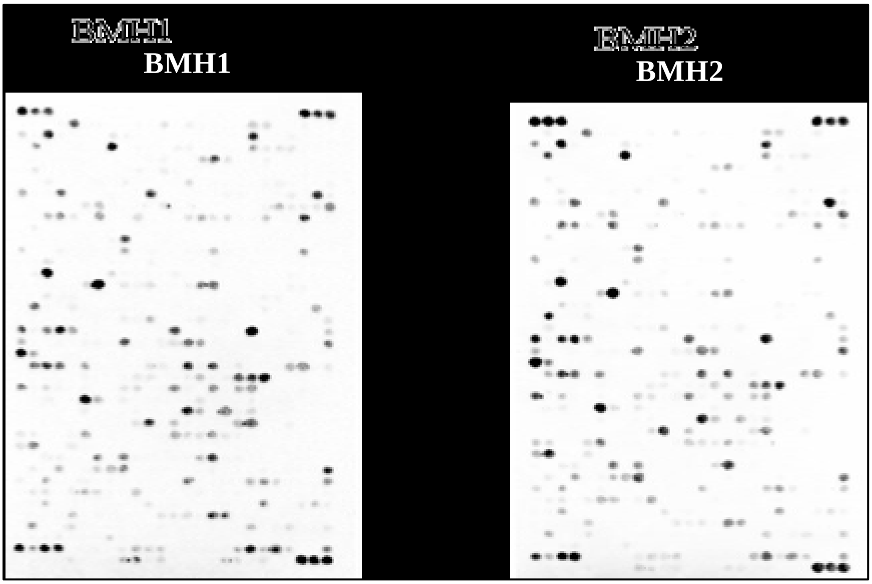


Figure 9 : PepSpot experiment where peptides derived from 13 ligand proteins are screened.

Summarizing we have identified sixty nine 14-3-3 binding peptides derived from thirteen known ligand proteins that constitute a set of experimental data that we can use as a test set for the selection criteria of our *in silico* screening. We also take into consideration, in a separate set, the remaining twenty fourth proteins stored in MINT as 14-3-3 partners although for these we do not have experimental information about the target peptide.

Filters benchmark against experimental data

At this stage our scope is to establish appropriate parameters regulating the stringency of each filter. We then test our system to identify the thresholds that lead to the largest recovery of experimentally determined partner and the smallest number of new inferred interactors of 14-3-3 domains from the full yeast proteome. Most of our filtering criteria do not have a set threshold, such as the regular expression, the peptide disorder prediction and the protein expression *in vivo* assessment. These three filters allow sorting of the peptides and proteins satisfying their requirements, and eliminate from further analysis the biological objects which do not meet their criteria. On the other hand the PSSM derived from mutagenesis, the comparative genomic filter and the common GO process criteria produce a quantitative output and suitable thresholds should be established to modulate their stringency.

Previous analysis in this work already gives us some elements to define filter thresholds. As shown in part1, our prediction of 14-3-3 partners has an accuracy ranging from 80 to 90% with PSSM thresholds from 0.48 to 0.50. Concerning the comparative genomic filter in conformity with the result obtained with ELM motifs and PROSITE patterns we require the regular expression to be conserved at least in 70% and ideally in 100% of the species. Regarding the GO process we need to explore if the increase of the depth (ranging from 1 to 8) of the common term in GO process tree increases the accuracy of the prediction. Finally we also generate two alternative sets of proteins co-localized with BMH1 and BMH2: the first includes all proteins (2.882) that are not separated by membranes from the nucleus and the cytoplasm, whereas the second includes only proteins (2.377) that are located in the nucleus or the cytoplasm.

Thus we explore a reasonable number of parameter combinations and run the 14-3-3 *in silico* screening on the full *S. cerevisae* proteome. We analyse the accuracy and coverage of the filters using as a reference set for validation the experimental data we have collected. Each filter is validated independently and in various combinations with the others. The accuracy and the coverage are calculated using the same formula of our previous studies:

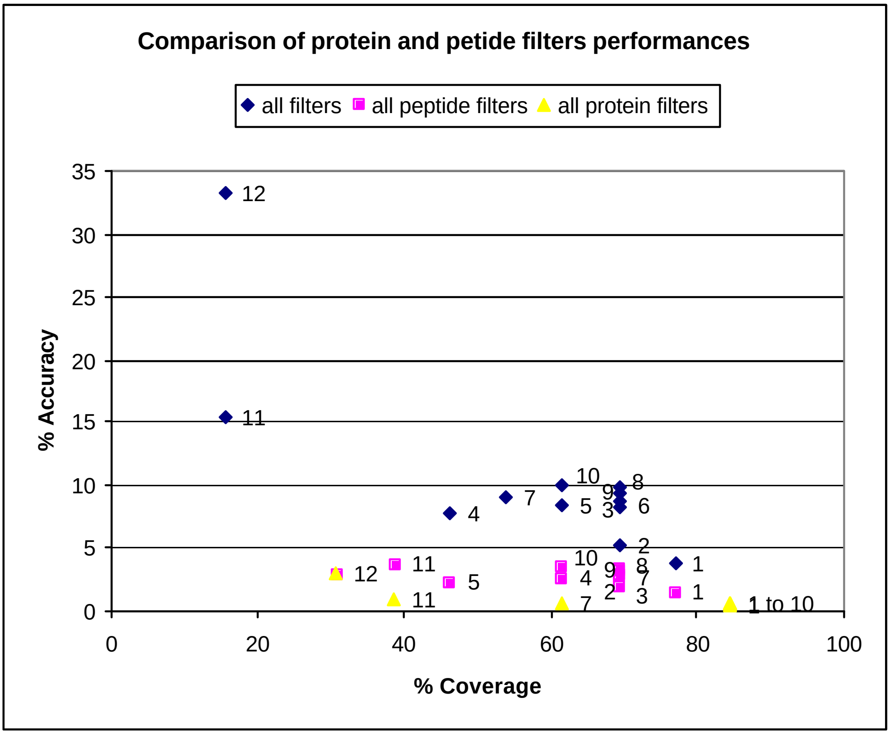
Accuracy = (True Positives / (True Positives + False Positives)) x 100
Coverage = (True Positives / Experimental Positives) x 100

In table 6 we report the 12 combinations of thresholds and the resulting accuracy and coverage of the full system calculated at peptide level (using the 69 high affinity peptides) and at protein level (referring to the 13 ligand proteins of our experimental set). The coverage and accuracy values we obtained are in general quite weak, but this is normal considering the discrepancy in size between the full yeast proteome that is analyzed (6.232 protein and the more than 250.000 peptides with a central serine) and our control dataset (13 ligand proteins and 69 binding peptides).

Set	PSSM score	% conserv Reg Exp Mut.	% conserv Reg Exp Phosp.	Co-located protein	GO term depth	Accuracy at protein level	Coverage at protein level	Accuracy at peptide level	Coverage at peptide level	N° recovered ligand vs 13	N° tot predicted ligand
1	0.47	60	60	2.882	1	3.84	76.92	3.00	14.49	10	250
2	0.48	70	70	2.882	1	5.29	69.23	4.97	14.49	9	161
3	0.48	70	80	2.377	2	8.25	69.23	8	14.49	9	100
4	0.48	80	80	2.377	2	8.42	61.53	8.25	14.49	8	87
5	0.48	70	90	2.377	2	7.69	46.15	7.77	13.04	6	72
6	0.48	70	87	2.377	2	8.73	69.23	8.47	10.14	9	94
7	0.48	70	87	2.377	3	9.09	53.84	8.79	14.49	7	70
8	0.49	70	87	2.377	2	9.27	69.23	9.09	14.49	9	88
9	0.50	70	87	2.377	2	9.89	69.23	9.80	14.49	9	82
10	0.51	70	87	2.377	2	10	61.53	10	13.04	8	72
11	0.52	90	100	2.377	4	15.38	15.38	20	4.34	2	11
12	0.52	100	100	2.377	5	33.33	15.38	37.5	4.34	2	4

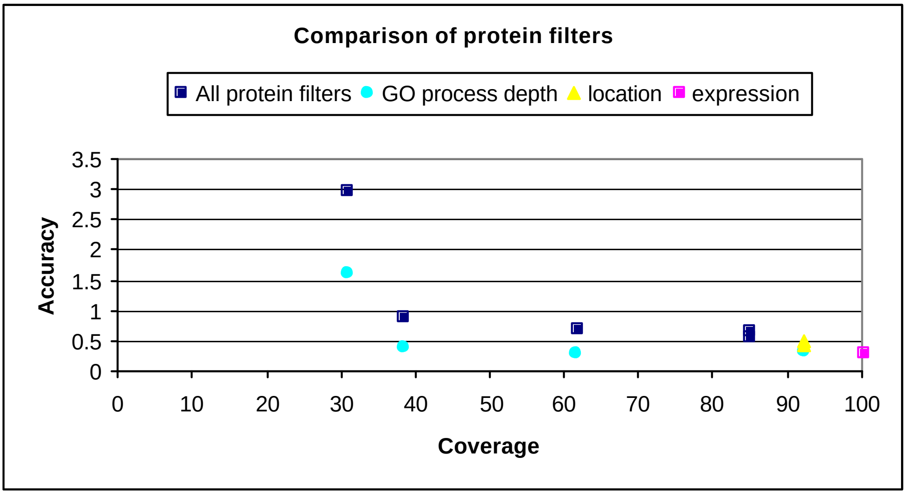
Table 6 : Accuracy and coverage of the full selection procedure using all filters. For the filters needing cut-off values we create different sets of threshold (first 5 columns) and report the performance of the resulting prediction on the corresponding line (last 6 columns). The ninth set of cut-off values (underscored line) id the one retained for our final prediction.

Surprisingly coverage calculated at peptide level is drastically lower than coverage at protein level. However we identify optimum parameters that lead to an overall prediction with almost 70% coverage and 10% accuracy.



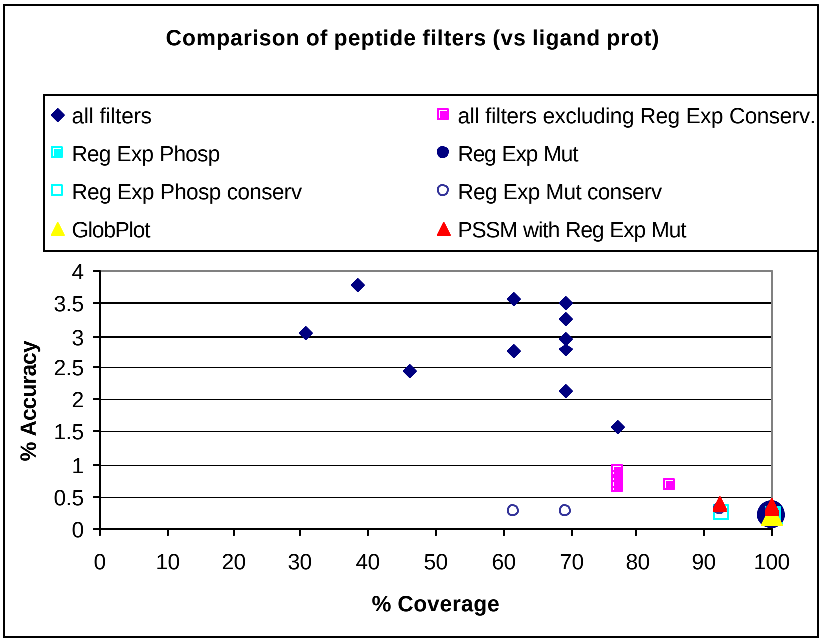
Graph 8 : Accuracy versus coverage plot of the prediction obtained using different sets of cut-off values (the thresholds set in each screening are reported in Table 6). The performance of the screenings with all filters is compared with those carried out considering either protein or peptide filters exclusively. Accuracy and coverage are calculated considering the 13 ligand proteins as reference set.

We first analyse the contribution of all protein filters and all peptide filters, combined in groups of two, to the full system accuracy (see Graph 8). As we expected peptide selection for 14-3-3 partners results on average more accurate than protein filtering criteria (Expression, co-localisation, and GO term sharing). However, when GO process term is shared at depth 5 (set n° 12 in Graph 8) protein filtering reaches the highest level of accuracy (see below Graph 9) but with considerable loss in coverage. Nevertheless, we can observe that protein criteria when applied with low cut-off values succeed in increasing the accuracy of 14-3-3 partners selection without affecting the prediction coverage (see for instance set n° 9 in Graph 8).

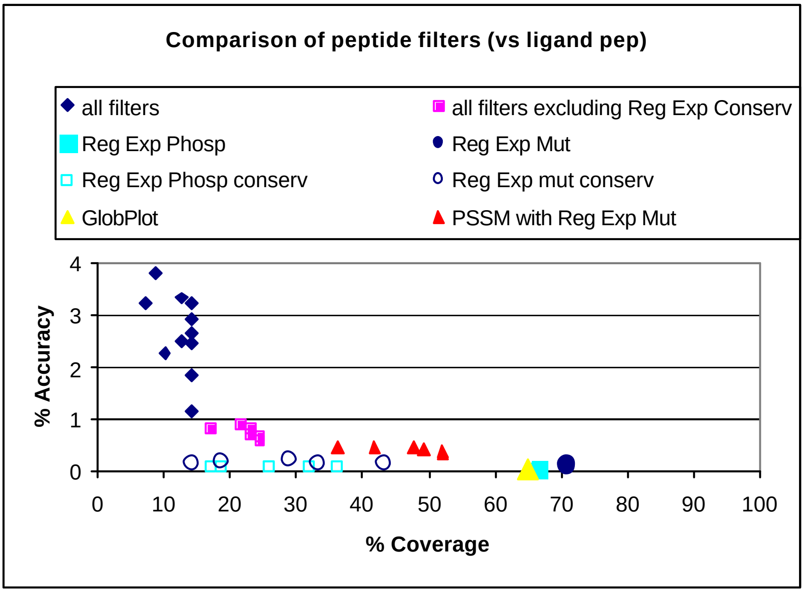


Graph 9 : Accuracy versus coverage plot of each protein filter independently and grouped as a whole at increasing values of GO term depth (from 1 to 5).

By analysing the accuracy of each protein selection criteria we can see that our expression filters, although rather unspecific, succeed in excluding spurious ORFs none of which is on our ligand set. The co-localisation criteria can be applied in the strictest manner by including only proteins that are exclusively located in the nucleus and the cytoplasm without significant variations in performances. Besides increasing the depth of the shared GO process term above level 2 strongly affects the proteins selection. We observe that depth 1 and 2 have identical performances whereas depth 3 reduces coverage without increasing accuracy as depth 4 and 5 do. So although GO criteria can increase accuracy when applied at high depth level, we choose depth_2 for our final prediction. Here again the combination of all the criteria i.e. the selection of the proteins satisfying all requirements, give the best results.



Graph 10 : Peptide filters accuracy versus coverage plot using the 13 known ligand protein as reference set.



Graph 11 : Peptide filters performances in recovering from the yeast genome the 69 known binding peptides.

Concerning the filters applied at peptide level, we report their performances at increasing cut-off values, both in terms of peptide and protein selection (see Graph 10 and Graph 11). It is interesting to see how these filters have a different coverage at peptide and protein level.

The regular expressions (derived from mutagenesis or collected for phosphorylation patterns) and the conservation according to comparative genomic filters seem to be a very unspecific criteria at protein level, whereas they are stringent at peptide level. Similarly the 14-3-3 PSSM and the regular expression derived from the mutagenesis data positively select around 40-50% of the peptides corresponding to 90% or more proteins. This coverage variation depends on the fact that very few peptides per protein are selected and that the small population of selected peptides identifies an almost equivalent number of proteins. In fact when we combine all the filters we find only one peptide that satisfies all requirements per protein.

In graph 11 we verify that the PSSM together with the regular expression derived from the mutagenesis is the highest accuracy filter in selecting 14-3-3 binding partner before the application of the other filters at the protein level. Moreover, we assess that the best PSSM score threshold is 0.50 when applied to full proteome screening. Furthermore we acknowledge that the comparative genomic filter validated around 50% of the regular expression matches and, when combined with other criteria, increases the overall accuracy values of a 4 folds with relatively little effect on coverage. In establishing the comparative genomic cut-off values (see Table 6) we determine that for the regular expression representing phosphorylation patterns the threshold can be set at 87% whereas the optimum cut-off for the regular expression to identify putative binding partners is 70%. This difference can be partially explained by the fact that most of the peptides match regular expressions from ELM phosphorylation pattern (see Table 3) that impose few restrictions (in three or fewer positions) compared to our regular expression for 14-3-3 target (9 positions defined). Thus it is not surprising that shorter stretches of sequences are better conserved than longer ones. This decision is also taken considering the results themselves (see below) and the peptides that would be excluded in case we use the same cut-off value for both applications of the comparative genomic filter.

Finally GlobPlot predictions have performances similar to the regular expressions, i.e. all proteins have at least one disordered peptide centred on a

Ser residue, but only 60% of our set of binding peptides are predicted to be unstructured. We observe that the experimental peptides already selected by all the other filters are also predicted to be in a disorder region and we consider this as an extra evidence of their biological relevance. Thus the GlobPlot filter results further supported the selection of putative 14-3-3 binding sites as it is unlikely that co-occurrence on the same peptide of all the requested sequence properties (14-3-3 binding, phosphorylation target, conservation in orthologous sequences and disorder propensity) happens by chance.

In silico prediction of 14-3-3 mediated interactions

When we apply the optimized set of cut-off values, the peptide filters select 302 out of the 250.000 decapeptides with a central serine in *S. cerevisae* proteome (0.12%). Beside all protein filters select 1.633 ORFs out of the initial 6.232 (26%) and the combination of both procedures leads to the prediction of 92 peptides belonging to 82 proteins (see Appendix). 9 proteins out of 13 from our experimental set are correctly predicted as putative ligand, whereas 4 proteins are discarded by the filtering system we set up. Moreover according to our prediction 4 proteins from the set of 24 known 14-3-3 ligands retrieved from the MINT database, for which we do not have experimental confirmation of the binding at peptide level, have an optimal binding site (see Table 7).

In the set of the 9 true positives ligands, 3 peptides belonging to GSY2, YAK1 and NTH1 are known to be phosphorylated *in vivo*, whereas the predicted binding sites for RTG3, MSN2 and REG1 are not known to be phosphorylated. The NTH1 peptide has the weakest conservation of the phosphorylation pattern but, because we have the *in vivo* evidence that this specific site is phosphorylated we decided to set the cut-off value accordingly. Moreover the validity of this specific peptide compared to other possible binding sites within NTH1 sequence is confirmed by BIACORE affinity measures (Panni *et al.* in prep.). Overall we also observe that in 5 cases out 9 the selected peptides are the best ligand site (rank first) within each protein sequence according to the ligand scanning PepSpot experiment.

Peptides	ORF	Pssm score	% Consv Reg Exp Mut	% Consv Reg Exp Phosp	GlobPlot prediction	ORF Location	GO term-Depth	ORF Expression (nolecules per cel)	Exp rank
True positive peptides from the experimental set									
SRIPFsERKLK	FAS1	0.52	100	100	disorder	cytoplasm	macromol metab-3	9.18E+04	1
VARPLsVPGSP	GSY2	0.53	100	100	disorder	cytoplasm,nucleus	polysac metab-5	1.46E+04	1
VRSPSsSFRAG	RTG3	0.56	90	100	disorder	cytoplasm,nucleus	metabolism-2	1.05E+03	2
GKQSSsLLSRL	FAS2	0.50	87.5	87.5	disorder	cytoplasm	macromol metab-3	1.70E+04	1
QRRLSsLSAFN	NTH1	0.53	71.43	100	disorder	cytoplasm	energy metab-5	1.84E+03	1
KSFKRsEHLKR	MSN2	0.52	100	100	disorder	cytoplasm,nucleus	metabolism-2	1.25E+02	6
LKRTRsMGLLD	REG1	0.52	100	100	disorder	cytoplasm	growth-2	2.56E+03	3
KRNFLsWKRGL	MYO2	0.51	88.89	88.89	disorder	bud, cytoplasm	growth-2	4.34E+03	8
KFRRAsLNSKT	YAK1	0.57	88.89	88.89	disorder	cytoplasm,nucleus	growth-2	obs GFP	1
True positive ligands from the MINT database									
LRGVKsLRFYL	KCS1	0.59	80	88.89	disorder	cytoplasm	growth-2	1.94E+02	NA
HVKKFsDFVSL	FUN31	0.50	77.78	88.89	disorder	cytoplasm	metabolism-2	9.67E+02	NA
KRHIRsVHSTE	MSN4	0.56	100	100	disorder	cytoplasm,nucleus	metabolism-2	obs GFP	NA
LMRAKsTKRIY	BOI2	0.50	100	100	disorder	bud, cytoplasm	growth-2	6.88E+02	NA
False negative peptides from the experimental set									
KLKSSsLLHLA	YPL110C	0.56	55.56	0	globular	cytoplasm	NA	2.30E+03	1
LKTSAsVRSRI	RTG2	0.39	100	100	Disorder	Cytoplasm	intrcel sign casc-4	3.26E+03	2
RLHNIsWRILN	MKS1	0.36	87.5	87.5	disorder	cytoplasm	metabolism-2	obs GFP	2
LFSAHsLPMDV	HEM15	0.68	100	NA	disorder	mitochondrion	metabolism-2	2.27E+03	5

Table 7 : Predicted 14-3-3 binding sites and their supporting experimental evidences. For each peptide we report the data that lead to its selection and its experimental rank among all peptides derived from the same protein in the PepSpot ligand scanning. In the lines with wrongly predicted peptides, we highlight with shadowed boxes the values causing its elimination.

Among the proteins known to be 14-3-3 ligands but not included in our ligand scanning experiment, we infer KCS1: this protein is one of the 3 most reliable partners (together with REG1 and NTH1) as it has two independent evidences of its interaction with BMH1 and BMH2 *in vivo* (see Table 5).

Moreover, we observe that the expression level of this set of ligands together with MSN2, GSY2, YAK1, and RGT3 is lower than 1000 molecules per cell, and this could explain why we did not detect them in our CoIp experiment.

In particular, YAK1 and MSN4 are suspected to have a very low number of copies as they have no detectable expression (Ghaemmaghami *et al.,* 2003 see section: 'Selection criteria at protein level') while proven to be existing gene product, as they were visualized as GFP fusion proteins (Huh *et al.,* 2003).

Concerning the false negatives, YPL110C is the only partner that we feel could be in fact a true negative. Apart from being an uncharacterized ORF whose GO process term is "unknown process", we cannot find a possible site both matching the patterns for 14-3-3 binding and showing some acceptable level of conservation of the relative regular expressions. We display in table 7 only one YPL110C peptide but this observation occurred along all its sequence. For RGT2 and MKS1 we report the peptides that satisfy almost all requirements except the score derived from our PSSM. Both these peptides have a limited number of positive charges at the N-terminus, together with hydrophobic amino acids in positions +1 and -1. This combination of residues is shown to be appropriate for 14-3-3 binding in the PepSpot experiment but is wrongly underrated by our PSSM score. HEM15 protein is excluded for two reasons: first it is annotated as a mitochondrial protein, second the best scoring peptide we report does not match any phosphorylation pattern. Regarding the localisation we do not discuss the validity of this assignment but only the static representation of proteins it gives. In fact as HEM15 is encoded by the nuclear genome and not by the mitochondrial, it is likely that the protein passes through the cytoplasm during its lifetime before reaching its final destination. Besides 14-3-3 proteins are known to be regulators of their partners location by retaining them either in cytoplasm or in the nucleus. Finally the HEM15 peptide belongs to a class of peptides having an XsLP pattern and showing good scoring in our PepSpot screenings. Despite this we could not find any phosphorylation regular expression matching such residues stretch and we do not observe any occurrence of this pattern among the known yeast phosphopeptides. On the other hand in mammalians cells at least seven partners of 14-3-3 isoforms are reported to have a phosphorylated binding site matching this XsLP pattern (Aitken *et al.,* 2002). Furthermore in high eukaryotes it has also been observed that some 14-3-3 complexes may not require phosphorylation of the target peptide. In our yeast system, we have evidence that the IPAWLsLPS template of our mutagenesis is a high affinity ligand of BMH1 and BMH2 that we choose in order to investigate 14-3-3 binding ability in absence of positive charges at the N-terminus of the site. For the time being we have no further clue to understand whether this binding consensus is discarded from our prediction because our collection of phosphorylation patterns is incomplete or because it is an artefact of our *in vitro* PepSpot data.

CONCLUSIONS

From this study case where we have compared experimental and *in silico* analysis of interaction mediated by 14-3-3 proteins interaction we can delineate general rules for protein interaction validation. We confirm that the search of coherent evidence supporting an interaction in different experimental data types such as proteins expression levels, proteins location in the cell and proteins functional characterisation is indeed a powerful mean to restrict the “social environment” wherein proteins can interact with each other. For co-expression analysis we take advantage of a proteomic study providing absolute concentration of each yeast protein. These data do not allow us to identify specific gene products that share the 14-3-3 expression pattern as RNA profiles could. On the other hand we circumvent possible regulation events at translation level and have direct information about the protein concentrations and not about their RNA precursor. Moreover, protein concentration data can sometimes explain discrepancies or the lack of overlap between the interaction networks detected by methods having different experimental resolution. In order to exploit protein location and function as validation criteria, GO is a very powerful tool for automatic annotation comparison. Although we could not use the GO compartment ontology and even for process we could not take advantage of the full tree depth, the growing usage of GO should quickly provide increasingly specific and high coverage mean for protein interaction validation. Each of these three criteria is rather unspecific in respect of interactions analysis but combining them helps to exclude artefact partners. In fact, although BMH1 and BMH2 are located in the two largest subcellular compartments (cytoplasm and nucleus) and do cooperate in several biological processes we could reduce by three quarters the number of yeast ORFs that satisfy all co-occurrence criteria under normal growth conditions. These elements are particularly valuable when coupled with sequence based analysis, *in vitro* screenings or even *in vivo* assays that can alter or overcome physiological conditions.

Besides we show how a similar multiple properties check of binding sites sequence can be used to infer or validate interacting partners. This approach requires appropriate investigations of the recognition determinants that rule partner recognition and therefore cannot be used for proteome wide interaction screening. Moreover it is applicable only when an interaction is mapped to binding region sequences on both partners Nevertheless we can already identify a number of domain-ligand pairs having characterized interaction features that are suitable for a sequence level validation similar to the one we proposed (see Table 1). Furthermore the systematic collection and analysis of such pairs of interacting sequence signature could also be very useful to predict the topology of complexes detected by co-immunoprecipitation.

Domain	**Domain PROSITE signature**	**Corresponding ligand ELM**
14-3-3	PS00796 ;PS00797	LIG_14-3-3_(1-3)
FHA	PS50006	LIG_FHA_1
GYF	PS50829	LIG_GYF
PDZ	PS50106	LIG_PDZ_(1-3)
SH2	PS50001	LIG_SH2_(1-7)
SH3	PS50002	LIG_SH3_(1-5)
WW	PS01159	LIG_WW_1-4
Dynein	PS01239	LIG_Dynein_DLC8_1
Clathrin	PS00224 PS00581	LIG_Clathr_ClatBox

Table 1: List of interaction template where both the domain signature and its ligands sequences requirements are well established and represented by either regular expression or position specific matrices.

We show that sequence conservation across several yeast species of 14-3-3 mediated interaction constrains successfully discriminates binding sites from spurious region matching by chance ligand consensus. Beltrao and Serrano, using a very similar approach showed that comparative proteomics can greatly increase the performance of a *consensus* based prediction of SH3 targets (Beltrao & Serrano submitted). However, binding sites conservation in close species is in contrast with co-variation of interacting sequence observed by Pazos and Valencia across bacterial proteomes (Pazos & Valencia 2002). Certainly both processes occur during evolution but further studies are required to determine an optimal divergence time for the genomes used when searching for conservation of ligand *consensus*. Moreover, it should be analyzed to what extent the co-variation of interacting sequences actually alters ligand *consensi* or domain signature conservation. From our preliminary studies we expect the ligand *consensi* to be more sensitive to co-evolution and potentially leading to different binding requirements in various species. On the other hand, we show that a regular expression representing the residues that must be avoided to allow the binding is highly conserved in the fifteen yeast species we considered. By extending our studied to more distant species for whom we have experimental data on 14-3-3 partners, we could verify whether our negative rules reflect general requirements of an extended conformation peptide to be lying on any 14-3-3 groove leaving enough flexibility to specific residues combination to occur.

Although the conservation of interaction is an open issue, currently many investigations are carried out to extend interaction information from one organism to another treating interaction data like functional or structural information to be extended by sequence similarity. The notion of "interologs" was first proposed by Walthout and co-workers (Walthout *et al.* 2000). Protein interaction networks determined in *S. cerevisae* are used to infer candidate interaction partners in *C. elegans*. By using BLAST reciprocal best hits methods they identify pairs of orthologs sequences that are known to interact in yeast and verify whether the corresponding interaction among worm "interologs" could be determined experimentally. Out of 216 inferred interaction 35 (16%) result to be true (Matthews *et al.* 2001). Interestingly they do not find detectable correlation between orthologs sequence similarity and likelihood of an interaction being conserved between yeast and worm. Furthermore, in this study the authors to circumvent ambiguities in defining orthologs with the best BLASTP reciprocal hits, finally extend interaction to all the good scoring homologues protein and state this do not affect the interolog prediction. This supports our decision to consider large orthology groups including in-paralogs as defined by INPARANOID software is indeed a correct approach when analysing interaction properties across proteomes.

Following similar approaches a human interaction map has been inferred from *S. cerevisae*, *C. elegans* and *D. megalonaster* experimental data (Lehner *et al.*, 2004). The inferred network is validated counting how many predicted interaction partner share a GO term. The authors find that considering interaction dataset for the single organism supported by at least two independent evidences or inferring human interaction only when two distinct organism networks support the same interolog pair are both very effective ways to improve the accuracy of the prediction. HomoMINT is the human interaction predictions regularly rebuilt on all the MINT data (including other mammalian model organisms) according to interologs method and available online as a parallel database (Cesareni *et al.*, 2005). Only 3.5% of the 17.000 predicted interactions in HomoMINT have some supporting experimental evidence stored in MINT or in any other interaction database.

Some answers come from investigations about transferability of protein interaction in relationship with the orthologs sequences similarity (Yu *et al.*, 2004). This study is carried out by considering in *H. pylori*, *S. cerevisae, C. elegans* and *D. megalonaster* experimental interaction networks (see Introduction Table1) and counting true positive interologs among these model organisms. It results that an interaction is very likely to be conserved in different organism if the joint sequence identity (geometric mean of the individual sequence identity) between the protein pair in the two organism is greater than 80%. This value correspond to BLASTP E-value lower than 10^{-10} for each candidate ortholog partner with 80% of the sequence length included in the alignments. Thus sequence identity criterion is in itself not extremely stringent (E-values ranged between 10^{-10} and 10^{-151}) but the fact that a large part of the protein must be involved confirm that the proteins overall domains structure should be maintained.

In this context the comparative genomic method we propose, based on the domain-ligand sequence signatures, could be used not only to validate a single interaction network but as a tool to infer interolog interactions across organisms.

APPENDIX

Below we report all the predicted 92 ligand peptides (belonging to 82 proteins) obtained running the *in silico* screening of 14-3-3 binding partners on the full *S. cerevisae* proteome with the optimized cut-off values. For practical reasons we do not report all criteria that are satisfied by these peptides, such as being disordered peptides and belonging to expressed proteins and being located in the nucleus or in the cytoplasm ORFs. Data are sorted according to the GO process term depth, shared by each ORF with BMH1 and BMH2.

Peptide	**ORF Name**	**PSSM score**	**% conserv Reg Exp Mut**	**% conserv Reg Exp Phosp**	**GO depth GO ID**
QRRLS**S**LSAFN	NTH1	0.53	71.43	100	5 GO:0006112
IARPL**S**VPGSP	GSY1	0.5224	100	100	5 GO:0005976
KKSVL**S**LANVG	GPH1	0.5051	100	100	5 GO:0005976
VARPL**S**VPGSP	GSY2	0.5383	100	100	5 GO:0005976
PVRVY**S**TPGDE	AAP1'	0.5011	100	100	5 GO:0005976
FQRAT**S**EARTA	PYC1	0.5501	85.71	100	4 GO:0005975
DFKSH**S**LPFAR	HAP5	0.7036	100	100	4 GO:0005975
VLIRW**S**LQHGY	YJR096W	0.5084	90	100	4 GO:0005975
FQRAT**S**EARTA	PYC2	0.5501	100	100	4 GO:0005975
TKFVR**S**LVREI	RPL36A	0.667	100	100	3 GO:0043170
SRIPF**S**ERKLK	FAS1	0.521	100	100	3 GO:0043170
VMRAI**S**LGLLK	RPN9	0.5986	100	100	3 GO:0043170
LRLAR**S**EKKFR	RPL13B	0.576	100	100	3 GO:0043170
GKQSS**S**LLSRL	FAS2	0.5006	87.5	87.5	3 GO:0043170
LRLAR**S**EKKFR	RPL13A	0.576	100	100	3 GO:0043170
KKLIQ**S**LPPTL	RPL38	0.6612	100	100	3 GO:0043170
HKFVK**S**SPVVP	ALA1	0.5851	77.78	100	3 GO:0043170
NSKSA**S**LFKQR	RPL24B	0.5691	100	100	3 GO:0043170
SSRNF**S**LAIID	PRE4	0.6318	100	100	3 GO:0043170
LKFVT**S**LPHRD	HOG1	0.6839	100	100	3 GO:0043170
SIRRR**S**FNVGS	GCN2	0.5386	100	100	3 GO:0043170
PTRHF**S**ALIGW	YNL045W	0.5193	100	100	3 GO:0043170

RRLVN**S**LKKDD	MRP8	0.5729	88.89	88.89	3 GO:0043170
RKMSK**S**LKNYP	ILS1	0.5669	100	100	3 GO:0043170
HTTTR**S**LRKIN	YTA7	0.5367	87.5	87.5	3 GO:0043170
RRFNS**S**IGRTA	RPL17B	0.5405	100	100	3 GO:0043170
KRASK**S**SGKMK	HIR3	0.5344	100	100	3 GO:0008151
KKAIR**S**ASTSA	MBP1	0.5048	88.89	88.89	3 GO:0008151
PLMRS**S**LFHNS	MBP1	0.5587	88.89	88.89	3 GO:0008151
FWKIF**S**SAKDH	RGA2	0.5355	80	100	3 GO:0008151
LLNPR**S**SFSGA	IDS2	0.5072	100	100	3 GO:0008151
TRVLN**S**LHLST	CDC26	0.5705	100	100	3 GO:0008151
HSKVC**S**LPTVC	SSL1	0.5946	100	100	3 GO:0008151
VNKSR**S**SGHFS	CDC53	0.5741	88.89	88.89	3 GO:0008151
GGRLS**S**KPIIM	SPT3	0.5137	87.5	100	3 GO:0008151
HRRSQ**S**ELTNL	SSD1	0.5163	77.78	100	3 GO:0008151
GGGRK**S**LFAPY	SSD1	0.5935	100	100	3 GO:0008151
VKSSA**S**LRWHS	RRD2	0.5105	100	100	3 GO:0008151
LKQPA**S**APVLP	HLR1	0.5734	100	100	3 GO:0008151
KQRRR**S**SYAFN	PAT1	0.5332	100	100	3 GO:0008151
IFSRF**S**TLFPN	MDM20	0.5228	100	90	3 GO:0008151
GPASF**S**LRSEN	RHC18	0.6273	87.5	87.5	3 GO:0008151
VKNPK**S**EFVVS	NGG1	0.5086	100	100	3 GO:0008151
RLRPF**S**YSKAD	RHO2	0.5209	100	100	3 GO:0008151
FKNRI**S**LNHSP	DBF20	0.5126	100	100	3 GO:0008151
TLRTS**S**PPFIP	DBF20	0.5144	100	100	3 GO:0008151
KKQTS**S**LKLAP	SLF1	0.5095	80	100	3 GO:0008151
KKVGF**S**AFGGL	MSH6	0.5268	87.5	87.5	3 GO:0008151
RRSFI**S**LRGSS	BCK1	0.548	100	88.89	3 GO:0008151
RAKMR**S**LFPFK	BEM3	0.6684	77.78	88.89	3 GO:0008151
HKVIN**S**LGVLD	ERB1	0.5105	90	90	3 GO:0008151
LKRTR**S**MGLLD	REG1	0.5275	100	100	3 GO:0008151
TRNNF**S**EHFKI	TOM1	0.5706	75	87.5	3 GO:0008151
LIKWK**S**LFPPF	DCC1	0.582	80	90	3 GO:0008151
LSASF**S**LKNGD	RPL6B	0.5794	100	100	3 GO:0008151
RKPMA**S**VPSCH	ADA2	0.5255	88.89	100	3 GO:0008151
LIVHL**S**SPLEG	LTE1	0.5534	75	100	3 GO:0008151
KRNFL**S**WKRGL	MYO2	0.5151	88.89	88.89	3 GO:0008151

KFRRA**S**LNSKT	YAK1	0.5743	88.89	88.89	3 GO:0008151
LSASF**S**LKNGD	RPL6A	0.5794	100	100	3 GO:0008151
YIIAK**S**SPSSI	PMD1	0.5662	100	100	2 GO:0030154
SRRSS**S**LAEYV	PMD1	0.5952	100	100	2 GO:0030154
VASRR**S**SHSTR	PMD1	0.5294	100	100	2 GO:0030154
RKSSS**S**DARRI	PMD1	0.5654	100	100	2 GO:0030154
FARGS**S**SPTLS	PMD1	0.6063	100	100	2 GO:0030154
LKFAS**S**SPISE	PMD1	0.6204	80	100	2 GO:0030154
ARTHR**S**SGKLP	YEL047C	0.5205	100	100	2 GO:0008152
STRYH**S**LHVNP	ABZ1	0.5439	100	100	2 GO:0008152
VQRFS**S**LTKPS	GPX2	0.5317	100	100	2 GO:0008152
KILLL**S**LKAGG	RAD5	0.5246	88.89	88.89	2 GO:0008152
VKRSK**S**DAASG	PAP1	0.5274	90	90	2 GO:0008152
VRSPS**S**SFRAG	RTG3	0.563	90	100	2 GO:0008152
DTRSF**S**SPQSD	SPT23	0.6391	100	100	2 GO:0008152
LPWRK**S**LNPKR	MET13	0.5076	90	90	2 GO:0008152
HVKKF**S**DFVSL	FUN31	0.5026	77.78	88.89	2 GO:0008152
LNVAK**S**LKIGG	ABD1	0.545	100	100	2 GO:0008152
TRSST**S**LRRRN	PBP1	0.6128	80	90	2 GO:0008152
QRPIF**S**TQFHP	CPA1	0.5236	100	100	2 GO:0008152
KRHIR**S**VHSTE	MSN4	0.5628	100	100	2 GO:0008152
RYKSC**S**AFAPI	YJL068C	0.5383	100	100	2 GO:0008152
VTRPL**S**LKTDI	GAT1	0.5326	87.5	87.5	2 GO:0008152
IGRPS**S**LHKAE	TRP5	0.536	100	100	2 GO:0008152
TCLRW**S**FPRDD	MET6	0.547	100	100	2 GO:0008152
LVNTA**S**LKRYM	ADE4	0.5471	88.89	88.89	2 GO:0008152
SISAR**S**SVHES	SRB8	0.5089	100	88.89	2 GO:0008152
IRLFR**S**ARRWV	ADE6	0.6121	87.5	87.5	2 GO:0008152
KSFKR**S**EHLKR	MSN2	0.5299	100	100	2 GO:0008152
KRHVR**S**VHSNE	MSN2	0.5639	100	100	2 GO:0008152
GFSAR**S**LRSLQ	TFC3	0.6489	77.78	100	2 GO:0008152
RRKLS**S**LSYEI	CET1	0.5629	88.89	88.89	2 GO:0008152
GKQVF**S**LLIKP	RPO31	0.5843	87.5	87.5	2 GO:0008152

MATERIALS AND METHODS

Writing this work I have tried to present methods and discuss results simultaneously. Because I did not personally carried out any wet experiment and used along my research a lot of bioinformatics software from public resources, my job basically consisted in a lot of python[1] () scripts that ensure the overall data flow. Thus here I just want to mention some relevant technicalities and expand some bioinformatics concepts that I give for granted in the text.

Gene Ontology

The Gene Ontology (GO) (Gene Ontology consortium, 2001) project was established to provide a common language to describe aspects of gene products biology. This project was started as a collaboration between three model organism databases, *Saccharomyces* Genome Database (SGD), FlyBase (Drosophila) and Mouse Genome Informatics (MGI); but nowadays a plethora of databases has adopted GO terms for gene product annotation (including UniProt, Genbank, ELM ect). The centralized creation of a comprehensive set of controlled vocabularies and their widespread usage, increase interoperability of many databases and facilitate annotation comparison of gene products across different species.

GO is a database of terms[2] for the description of the **Molecular function**, the **Biological process** and the **Cellular component** associated to gene products. Molecular function is defined as the biochemical activity of a gene product including specific binding to ligands (e.g. ‘enzyme’, ‘trasporters’, ’receptor’). Biological process refers to a biological objective to which the gene or proteins contributes to (e.g. ‘cell growth’, ‘translation’, ’pyrimidine metabolism’). Cellular component refers to the place in the cell where a gene product is active (e.g. ‘nucleus’, ‘vacuole’, ‘Golgi apparatus’). Molecular function, biological process and cellular component are three independent attributes of gene products and thus are maintained as three separate ontologies.

[1] *http://www.python.org/*

[2] *http://www.geneontology.org/*

An ontology does not provide only a list of controlled vocabularies but it also represents the relationship among its terms through a hierarchical tree. More precisely, GO terms are organized in a directed acyclic graphs (DAG) where higher level terms can have many children terms and also each child term can have more than one parent. Moreover, within an ontology terms can have two different relationships, a child can be "part of" or "instance" of its parents (see Figure 1). For example, in the cellular component ontology, the vacuole is part of the cytoplasm and also is an instance intracellular membrane bound organelle.

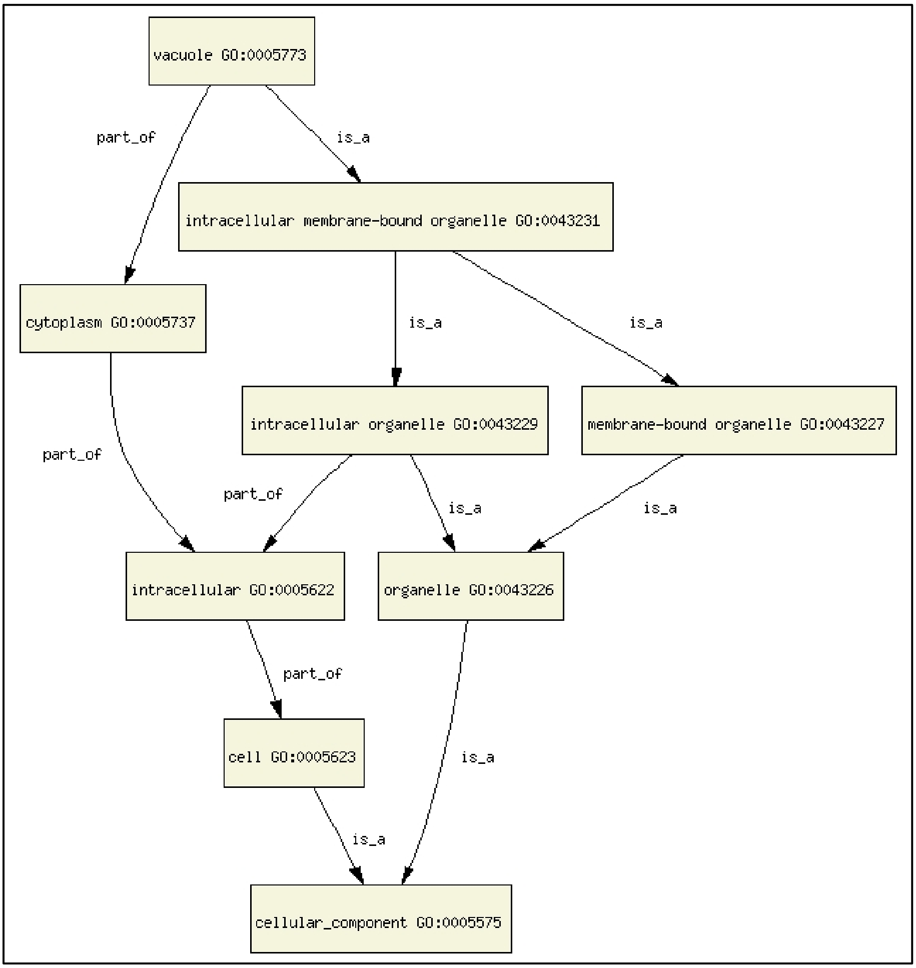


Figure 1: Fragment of the cellular component ontology graphically visualized with AMIGO browser (http://www.godatabase.org/cgi-bin/amigo).

GO database itself does not store gene products, but only keeps the ontologies updated to the scientific community needs, curating each term definition and maintaining the vocabulary trees.

The assignment of GO terms is a work performed by the single databases that refer to GO for the annotation of the gene products they store[1]. For instance, SGD is responsible for the annotation of the *S. cerevisae* ORFs according to GO terms and GOA (GO Annotation initiative) aims at adding GO cross references within UNIPROT entries starting from the model organisms. GO user databases have common rules for assigning detailed terms based on the evidences that support the association of a given gene product to a specific term. Evidences are reported with a three letter code (e.g. IDA stands for 'Inferred from Direct Assay' or ISS refers to 'Inferred from Sequence Similarity') and attached to every assignment. In this way end users can select subsets of gene products annotations according to their own idea of information reliability.

Combining data from GO database and the related gene product annotation efforts one can build very powerful analysis tool. For example the GO Term Finder at SGD[2] can be queried with any set of *S. cerevisae* ORFs and it returns the GO terms that are shared by the ORFs of interest. ORFs may not have identical GO terms assigned but navigating upwards the ontology three a common parent term can identified, providing clues about what the ORFs have in common in their biology. Alternatively many tools are built to analyse experimental results according to the GO terms. One of these is the hierarchView embedded in IntAct database that allows the exploration of interaction networks in the context of the GO annotations assigned to interacting proteins.

Curation of protein-protein interaction

During my PhD years I have been deeply involved in the development of MINT (Zanzoni, A. *et al.*, 2002) protein-protein interaction database. Consequently I participated in the Proteomics Standards Initiative (PSI) (Hermjakob *et al.*, 2004a) that aims at establishing common standards in

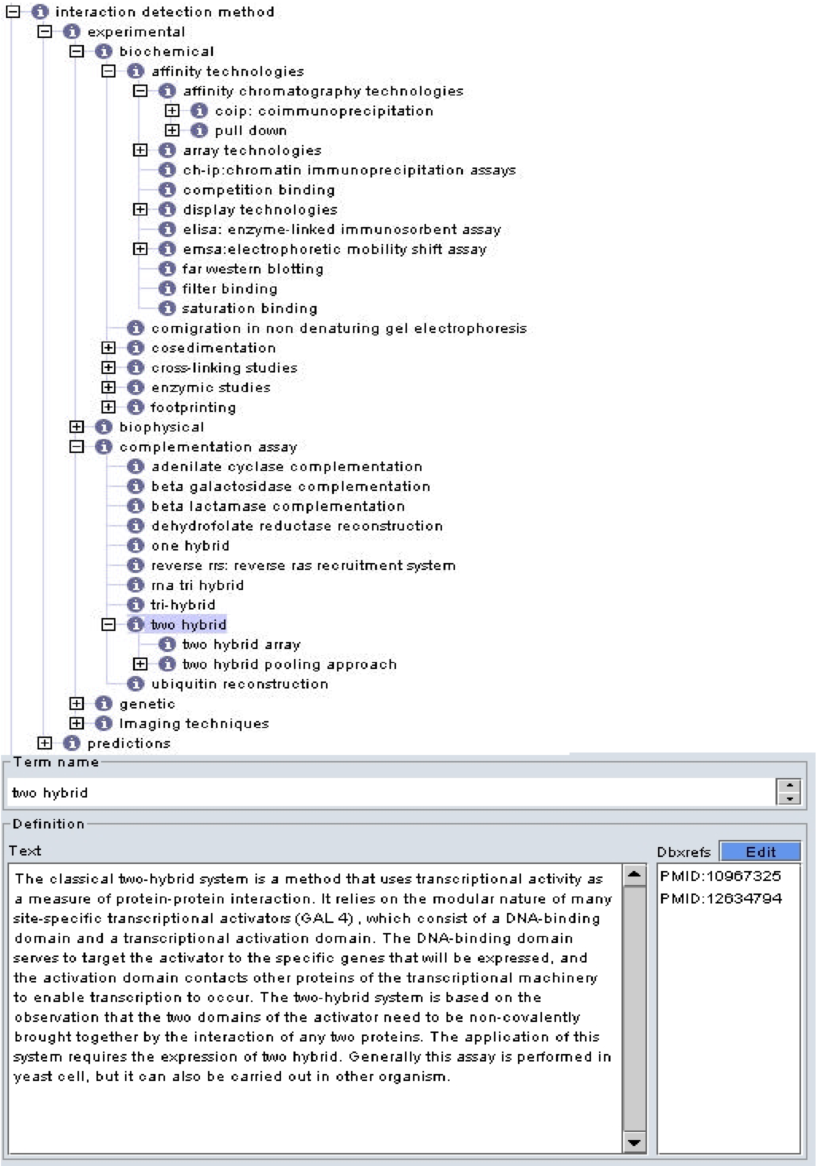
[1] *http://www.geneontology.org/GO.current.annotations.shtml*

[2] *http://db.yeastgenome.org/cgi-bin/GO/goTermFinder*

order to exchange data among protein interaction databases: MINT, Intact (Hermjakob *et al.,* 2004b), DIP (Xenarios *et al.,* 2002), CYGD (Guldener U *et al.,* 2005), BIND (Alfarano *et al.,* 2005) and STRING (von Mering, C. *et al.,* 2005).

My main contribution to the protein databases consisted in the systematic curation of most of the high through put interaction sets mentioned in this work, including of course the PepSpot data coming from Prof. G. Cesareni laboratory (more than 40.000 interactions in total). Protein interaction curation is a process of extraction and formalisation of experimental results published in the literature, in order to submit and store it into a dedicated database. The first step is the assignment of a UNIPROT protein sequence database (Bairoch *et al.,* 2005) reference to the set of interacting proteins. This is a critical point, because most times the authors of a paper provide a list of protein or gene names that may not be univocally related to a single protein sequence. Because protein names are not standardized, it is absolutely necessary to associate interaction to specific sequences in order to integrate information from different sources. Thus a specific procedure has been established to assign, in a reliable manner sequence references to interacting proteins. This could also lead to discard interactions because an appropr iate reference could not be found. The second step of the submission is the extraction of any protein specific feature known to be required for the interaction to occur. The so called “interactor features” include the information, if present, about the bin ding domain responsible for the interaction, the post translation modification and mutations required for the interaction to occur. The third step, consists in the annotation of the interaction itself, in term of what kind of interaction it is (aggregation, or an enzymatic reaction ect..), where it takes place (the organisms, the cell line, and if known the cellular compartment), kinetic information if provided. The fifth and last step is the description of the experimental evidence supporting the interaction.

Practically all databases in their submission process take advantage of controlled vocabularies presenting to curators alternative terms to describe the object being annotated. For instance, drop down menu are still in use to list all possible post translation modifications or experimental methods. I was deeply involved in the transformation of all these drop down menus, from the consorted protein interaction databases, into a single formal ontology that could be used as common reference glossary to describe protein interactions.



- interaction detection method
  - experimental
    - biochemical
      - affinity technologies
        - affinity chromatography technologies
          - coip: coimmunoprecipitation
          - pull down
        - array technologies
        - ch-ip:chromatin immunoprecipitation assays
        - competition binding
        - display technologies
        - elisa: enzyme-linked immunosorbent assay
        - emsa:electrophoretic mobility shift assay
        - far western blotting
        - filter binding
        - saturation binding
      - comigration in non denaturing gel electrophoresis
      - cosedimentation
      - cross-linking studies
      - enzymic studies
      - footprinting
    - biophysical
    - complementation assay
      - adenilate cyclase complementation
      - beta galactosidase complementation
      - beta lactamase complementation
      - dehydrofolate reductase reconstruction
      - one hybrid
      - reverse rrs: reverse ras recruitment system
      - rna tri hybrid
      - tri-hybrid
      - two hybrid
        - two hybrid array
        - two hybrid pooling approach
      - ubiquitin reconstruction
    - genetic
    - Imaging techniques
  - predictions

Term name

two hybrid

Definition

Text

The classical two-hybrid system is a method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature of many site-specific transcriptional activators (GAL 4) , which consist of a DNA-binding domain and a transcriptional activation domain. The DNA-binding domain serves to target the activator to the specific genes that will be expressed, and the activation domain contacts other proteins of the transcriptional machinery to enable transcription to occur. The two-hybrid system is based on the observation that the two domains of the activator need to be non-covalently brought together by the interaction of any two proteins. The application of this system requires the expression of two hybrid. Generally this assay is performed in yeast cell, but it can also be carried out in other organism.

Dbxrefs

PMID:10967325
PMID:12634794

Figure 2 : Subpart of the PSI ontology tree describing the interaction detection method and an example of terms definition. This hierarchical representation can be visualized with the DAG editor tool (see http://www.geneontology.org/GO.tools.html).

Following the Gene Ontology example, we provided a definition and references for every term and built a hierarchical tree representing the relationship among terms (see Figure 2). The resulting PSI ontology ensures that all databases give the same meaning to the terms employed and the hierarchy facilitates the analysis, according to a single classification, of interaction data coming from different databases. This ontology is nowadays dynamically maintained to include terms for new methods and also to expand PSI efforts to protein interaction in a broader sense, including protein-nucleic acid and protein-small molecules interactions.

Building a proteomic table to combine *S. Cerevisae* data.

In many occasions while achieving this work we integrate experimental results and also Gene Ontology annotation with sequence features. All these kinds of information were associated to *S. cerevisae* ORFs via different pointers such as their gene names, their UniProt accession number or simply their sequence. Thus we needed to create a reference table to navigate univocally from any ORF or gene name to its protein sequence or UniProt accession number. Fortunately *S. cerevisae* is a well characterized model organism, which genome and proteome are rather simply related and where every ORF has a unique systematic name (e.g. YER177W). We built our *S. cerevisae* proteomic table taking the non-redundant set of protein sequence from Integrat8[1] and merged it with the gene name registry table from SGD fungal database[2] (Balakrishnan *et al.*, 2005) using ORFs systematic names as key element. The resulting table lists 6200 gene product one per row, and in every row there is the systematic ORF name, the gene name (e.g. BMH1), the UniProt primary and secondary accession number (e.g. P29311) and the full amino acid sequence. Using this table, whatever reference was used to point to a *S. cerevisae* protein, we were able to map it and associate information from different sources to the same biological object.

[1] ftp://ftp.ebi.ac.uk/pub/databases/integr8/fasta_files/proteomes/4932.FASTAC.gz

[2] ftp://ftp.yeastgenome.org/yeast/data_download/gene_registry/registry.genenames.tab

Protein functional patterns

A functional pattern or motif is a representation of a particular cluster of residues characteristic for a particular biological function. Although pattern and motif are almost synonym, in this work we preferentially use the term pattern to refer to a long and variable functional signature whereas we call motif simpler and shorter patterns. Functional motifs are discovered either by experimental evidence that allows identification of residue directly responsible for the function or by sequences comparison of proteins sharing the same function, or by combining both information. The resulting motif can be used to infer the function of proteins that do not show globally in their sequences strong similarity with any other protein of known function. The fact that such sequences have a given cluster of residues in fixed positions is considered a reasonable evidence to predict their function. The accuracy of function inference depends on the "reliability" of a pattern: a pattern is considered "reliable" if most of the proteins (ideally all) sharing the function match the pattern (the so called true positives) and very few proteins (ideally none) outside the family match the pattern (the so called false positives).

Functional motifs can have different "grammar" for the representation of their sequence signature. Important residues can be represented as columns of a multiple alignment, a regular expression, a position specific scoring matrix (PSSM) (Gribskov *et al.,* 1987) or a hidden Markov model (HMM) (Baldi *et al.,* 1994). Each type of pattern "grammar" requires specific algorithms to search matches along sequences and returns different results.

Regular expressions are deterministic representations whereas PSSM and HMM are probabilistic. Regular expression patterns are deterministic in the sense that a sequence either has or does not have a match. Whereas PSSM and HMM are probabilistic because they assign a score to any sequence match related to its probability of being a biologically meaningful occurrence of the pattern.

A regular expression can be formally defined as a "grammar" to search patterns. It allows quick searching of substrings (i.e. a pattern match) within another string (i.e. a protein sequence). Regular expressions are made of normal characters and metacharacters that have special meanings.

Rule or metacharacters	**PROSITE standard**	**ELM standard**
Amino acids representation	The standard IUPAC one-letter code	The standard IUPAC one-letter code
Representation of a mandatory residue in a position	The corresponding one character	The corresponding one character
Metacharacter used to describe wildcard specifying that any amino acid is allowed in a given positions	**‘x’** (lower case x)	**‘.’** (period)
Metacharacter used to describe a set of allowed residues in a given position.	Square brackets, for example **[KR]** stands for Lys or Arg.	Square brackets, for example **[KR]** stands for Lys or Arg.
Metacharacter used to denote a set of residues that are <u>not accepted</u> in a given position.	Curly brackets, for example **{DE}** stands for any amino acid except Asp or Glu.	Square brackets with a caret, **[^DE]** stands for any amino acid except Asp or Glu.
Metacharacter used to describe separate positions	**‘-‘** (hyphens)	none
Metacharacter used to describe repetition of an element for an exact number of *n* times.	**(n)** for example A(4) stands for a repetition of four Ala	**{n}** for example A{4} stands for a repetition of four Ala
Metacharacter used to describe repetition of an element for a number of times ranging from *n* to *m*	**(n,m)** for example A(1,3) stands for a repeat of one, or two or three Ala in consecutive positions	**{n,m}** for example A{1,3} stands for a repeat of one, or two or three Ala in consecutive positions
Metacharacter used to restrict a pattern starting at the N - terminal	**‘<’**	**‘^’**
Metacharacter used to restrict a pattern ends at the C-terminal	**‘>’**	**‘$’**
Metacharacter used to describe the end of the pattern	**‘.’** (period)	none

Table 1 : Syntax for regular expressions according to PROSITE and ELM standards

Metacharacters are not mandatory, but are very useful to provide flexibility to a regular expression. For instance they are required to specify that some positions can contain a given subset of amino acids or even any kind of amino acid. A regular expression containing only normal characters is the most stringent type and implicates that the substring that we are searching must be identical to the string used as regular expression in every position. The syntax adopted for PROSITE pattern (Hulo *et al.,* 2004) and its equivalency with ELM is reported in table 1.

In this work we follow ELM (Puntervoll *et al.,* 2003) standard that is also the syntax dictated by python programming language. Thus, when we used PROSITE patterns we automatically convert their syntax according to this equivalency table. A typical Position Specific Scoring Matrix (PSSM) has 20 columns and as many rows as there are positions in the pattern. It generally stores observed frequencies of residues either resulting from experiment or from analysis of multiple alignments. The observed frequencies of residues (that may not be detected for all possible residues) are often combined with substitution matrices (Dayhoff *et al.,* 1972) to provide a final weight to any possible amino acid in each position. A substitution matrix is a 20x20 table where a value is associated to each possible residue-residue pair. Pairs of identical residues are associated to the highest values while the lowest values are associated to pairs of residues that have very different chemico-physical properties. The resulting PSSM, when used to assign function to an unknown protein are able to detect weaker similarities compared to regular expressions because even new combination of residues within a pattern can have a good score.

HMM is a dynamic kind of statistical profile. Like an ordinary profile, it is built by analyzing the distribution of amino acids in a training set of a protein family and can be used to perform sensitive database searching using the statistical description of a sequence family consensus. However, an HMM cannot be built on short sequences like the peptides resulting from PepSpot experiments, this is why this method have been discarded in this work.

Currently, there are many sequence pattern databases that describe protein globular domains, including PFAM (Bateman *et al.,* 2002), SMART (Letunic *et al.*, 2004), PROSITE (Hulo *et al.,* 2004), PRODOM (Bru *et al.,* 2005) and BLOCKS (Henikoff *et al.,* 2000), each one using different discovery methods and different pattern representations. The inferred protein domains from all these database are also stored and harmonized under the

INTERPRO umbrella (Mulder *et al.,* 2005). These resources provide an accurate overview of the domain architecture of a polypeptide and can therefore help to infer the functions of uncharacterized proteins. ELM is the only database that focuses on un-structured short sequence patterns that are required for post-translation modification, protein interaction or cell compartment targeting. In a very schematic way, one can say that the linear motifs stored in ELM are the targets of the protein domains whose pattern representations are collected in the above mentioned databases.

REFERENCES[1]

- **Aitken A,** Baxter H, Dubois T, Clokie S, Mackie S, Mitchell K, *et al.* Specificity of 14-3-3 isoform dimer interactions and phosphorylation. Biochem Soc Trans. **2002** 30:351-60.
- **Alfarano C**, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, *et al*. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. **2005** 33:D418-24.
- **Bader GD, Hogue CW**. Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol. **2002** 20:991-7.
- **Bairoch A**, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, *et al.* The Universal Protein Resource (UniProt). Nucleic Acids Res. **2005** 33:D154-9
- **Balakrishnan R**, Christie KR, Costanzo MC, Dolinski K, Dwight SS, Engel SR *et al.* Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the Saccharomyces Genome Database (SGD). Nucleic Acids Res. **2005** Jan 33:D374-7. see http://www.yeastgenome.org/)
- **Baldi P**, Chauvin Y, HunkApillar T, McClure M. HiddenMarkov models of biological primary sequence information. Proc. Natl. Acad. Sci. USA. **1994** 91:1059-1063.
- **Bateman A**, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, *et al.* The Pfam protein families database. Nucleic Acids Res. **2002** 30:276-80. see http://www.sanger.ac.uk/Software/Pfam/
- **Beck T**, Hall MN. The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. Nature. **1999** 402:689-92
- **Beltrao P, Serrano L**. Efficient comparative genomics and disorder prediction identify biologically relevant protein interactions. *Submitted*
- **Bouwmeester T**, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, *et al.* A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat Cell Biol. **2004** 6:97-105.

[1] When a paper has more than six authors the abbreviation “*et al.,*” is used

- **Brazma A**, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. **2001** 29:365-71
- **Brinkworth RI**, Breinl RA, Kobe B. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. Proc Natl Acad Sci U S A. **2003** 100:74-9.
- **Bru C**, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. **2005** Jan 33:D212-5. see http://prodes.toulouse.inra.fr/prodom/current/html/home.php
- **Bruckmann A,** Steensma HY, Teixeira De Mattos MJ, Van Heusden GP. Regulation of transcription by Saccharomyces cerevisiae 14-3-3 proteins. Biochem J. **2004** 382:867-75
- **Cesareni G**, Ceol A, Gavrila C, Montecchi-Palazzi L, Persico M, Schneider MV. Comparative interactomics. FEBS Lett. **2005** *in press*
- **Cherry JM**, Ball C, Weng S, Juvik G, Schmidt R, Adler C, *et al.* Genetic and physical maps of Saccharomyces cerevisiae. Nature. 1997 387:67-73.
- **Cliften P**, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, *et al.* Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science. **2003** 301:71-6.
- **Dandekar T**, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. **1998** 23:324-8
- **Dayhoff MO,** Eck RV Park CM. A model of evolutionary change in proteins. Natl. Biomed. Res. Found. Washington, DC. **1972** 5:89.
- **Deane CM**, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics. **2002** 1:349-356.
- **Dietrich FS**, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, *et al.* The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science. **2004** 304:304-7.
- **Dujon B**, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, *et al.* Genome evolution in yeasts. Nature. **2004** 430:35-44.
- **Felsenstein, J. 2004**. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome*

Sciences, University of Washington, Seattle. http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html

- **Ficarro SB**, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, *et al.* Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol. **2002** 20:301-5.
- **Fields S, Song O**. A novel genetic system to detect protein-protein interactions. Nature. **1989** 340:245-6
- **Fitch WM.** Distinguishing homologous from analogous proteins. Syst Zool. **1970** 19:99-113.
- **Frank R.** Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. Tetrahedron **1992** 48:9217-9232
- **Gaillardin C**, Duchateau-Nguyen G, Tekaia F, Llorente B, Casaregola S, Toffano-Nioche *et al.* Genomic exploration of the hemiascomycetous yeasts: 21. Comparative functional classification of genes. FEBS Lett. **2000** 487:134-49
- **Galagan JE,** Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, *et al.* The genome sequence of the filamentous fungus Neurospora crassa. Nature. **2003** 422:859-68.
- **Gavin AC**, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. **2002** 415:141-7
- **Ge H,** Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet. **2001** 29:482-6.
- **Gene Ontology Consortium,** Creating the gene ontology resource: design and implementation. Genome Res. **2001** 11:1425-33
- **Ghaemmaghami S,** Huh WK, Bower K, Howson RW, Belle A, Dephoure N, *et al.* Global analysis of protein expression in yeast. Nature. **2003** 425:737-41.
- **Giaever G,** Chu AM, Ni L, Connelly C, Riles L, Veronneau S, *et al.* Functional profiling of the Saccharomyces cerevisiae genome. Nature. **2002** 418:387-91
- **Gribskov M,** McLachlan AD, Eisenberg D. Profile analysis:Detection of distantly related proteins. Proc. Natl. Acad. Sci. USA. **1987** 84:4355-4358.

- **Guldener U,** Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, *et al.* CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res. **2005** 33:D364-8.
- Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. **2004** 430:88-93.
- **Henikoff JG,** Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. **2000** 28:228-30. see http://blocks.fhcrc.org/blocks/
- **Hermjakob H**, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, *et al.* The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol. **2004a** 22:177-83.
- **Hermjakob H,** Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, *et al.* IntAct: an open source molecular interaction database. Nucleic Acids Res. **2004b** 32:D452-5.
- **Ho Y**, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. **2002** 415:180-3.
- **Huh WK**, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, *et al.*. Global analysis of protein localization in budding yeast. Nature. **2003** 425:686-91
- **Hulo N**, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, *et al.* Recent improvements to the PROSITE database. Nucleic Acids Res. **2004** 32:D134-7. see http://www.expasy.org/prosite/
- **Huynen MA**, Snel B, von Mering C, Bork P. Function prediction and protein networks. Curr Opin Cell Biol. **2003** 15:191-8.
- **Ito T**, Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome Proc. Natl. Acad. Sci. USA **2001** 98:4569-4574,.
- **Ito T**, Ota K, Kubota H, Yamaguchi Y, Chiba T, Sakuraba K, *et al.*. Roles for the two-hybrid system in exploration of the yeast protein interactome. Mol Cell Proteomics. **2002** 1:561-6.
- **Jansen R**, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res. **2002** 12:37-46.

- **Jin J**, Smith FD, Stark C, Wells CD, Fawcett JP, Kulkarni S, *et al*.. Proteomic, functional, and domain-based analysis of *in vivo* 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. Curr Biol. **2004** 14:1436-50.
- **Jones T**, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, *et al*. The diploid genome sequence of Candida albicans. Proc Natl Acad Sci U S A. **2004** 101:7329-34
- **Kellis M**, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. **2004** 428:617-24.
- **Kellis M**, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. **2003** 423:241-54.
- **Kemmeren P**, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, *et al*. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol Cell. **2002** 9:1133-43.
- **Kersey P**, Bower L, Morris L, Horne A, Petryszak R, Kanz C, *et al*. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res. **2005** 33:D297-302. see http://www.ebi.ac.uk/integr8
- **Kramer A,** Schneider-Mergener J (1998) Synthesis and screening of peptide libraries on continuous cellulose membrane supports. Methods Mol Biol 87:25-39.
- **Landgraf C**, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, *et al*. Protein interaction networks by proteome peptide scanning. PLoS Biol. **2004** 2:E14.
- **Lehner B**, Fraser AG. A first-draft human protein-interaction map. Genome Biol. **2004** 5:R6
- **Letunic I,** Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, *et al*. SMART 4.0: towards genomic data integration. Nucleic Acids Res. **2004** 32:D142-4. see http://smart.embl-heidelberg.de/
- **Li S,** Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M *et al*. A map of the interactome network of the metazoan C. elegans. Science. **2004** 303:540-3.
- **Linding R,** Russell RB, Neduva V, Gibson TJ (**2003**) GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31:3701-3708. see http://globplot.embl.de/

- **Liu D,** Bienkowska J, Petosa C, Collier RJ, Fu H, Liddington R. Crystal structure of the zeta isoform of the 14-3-3 protein. Nature. **1995** 376:191-4.
- **Mackintosh C.** Dynamic interactions between 14-3-3 proteins and phosphoproteins regulate diverse cellular processes. Biochem J. **2004** 381:329-42
- **Marcotte EM,** Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science. **1999** 285:751-3.
- **Matthews LR,** Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res. **2001** 11:2120-6.
- **Mayordomo I,** Regelmann J, Horak J, Sanz P. Saccharomyces cerevisiae 14-3-3 proteins Bmh1 and Bmh2 participate in the process of catabolite inactivation of maltose permease. FEBS Lett. **2003** 544:160-4.
- **Moriya H,** Shimizu-Yoshida Y, Omori A, Iwashita S, Katoh M, *et al.* A DYRK family kinase, translocates to the nucleus and phosphorylates yeast Pop2p in response to a glucose signal. Genes Dev. **2001** 15:1217-28.
- **Mulder NJ,** Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, *et al.* InterPro, progress and status in **2005**. Nucleic Acids Res. **2005** 33:D201-5. see http://www.ebi.ac.uk/interpro/index.html
- **Obenauer JC,** Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. **2003** 31:3635-41. see http://scansite.mit.edu/
- **Orchard S,** Taylor CF, Hermjakob H, Weimin-Zhu, Julian RK Jr, Apweiler R. Advances in the development of common interchange standards for proteomic data. Proteomics. **2004** 4:2363-5.
- **Panni S,** Montecchi-Palazzi L, Cesareni G. Binding specificity studies of yeast 14-3-3 combining SPOT synthesis data with comparative genomic analysis. *Manuscript in preparation*
- **Pawson T, Nash P** (**2003**) Assembly of cell regulatory systems through protein interaction domains. Science 300: 445-452.
- **Pazos F, Valencia A.** In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins. **2002** 47:219-27.
- **Pellegrini M,** Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis:

protein phylogenetic profiles. Proc Natl Acad Sci U S A. **1999** 96:4285-8.

- **Pellegrini M.** Computational methods for protein function analysis. Curr Opin Chem Biol. **2001** 5:46-50.
- **Peri S,** Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, *et al.* Nucleic Acids Res. **2004** 32:D497-501 see http://www.hprd.org
- **Puntervoll P,** Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, *et al.* ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. (**2003**). Nucleic Acids Res. 31: 3625-3630. see http://elm.eu.org/
- **Rain JC,** Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, *et al.* The protein-protein interaction map of Helicobacter pylori. Nature. **2001** 409:211-5.
- **Remm M**, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. **2001** 314:1041-52
- **Rice P,** Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **2000** 16:276-7 see http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/
- **Rubio MP,** Geraghty KM, Wong BH, Wood NT, Campbell DG, Morrice N, *et al.* 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking. Biochem J. **2004** 379:395-408.
- **Scott JK, Smith GP**. Searching for peptide ligands with an epitope library. Science. **1990** 249:386-390
- **Souciet J**, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, *et al.* Genomic exploration of the hemiascomycetous yeasts: a set of yeast species for molecular evolution studies. FEBS Lett. **2000** 487:3-12.
- **Sprinzak E**, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? J Mol Biol. **2003** 327:919-23.
- **The Gene Ontology Consortium**: Creating the gene ontology resource: design and implementation. Genome Res **2001** 11:1425-1433 .
- **Tong AH**, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, *et al.* A combined experimental and computational strategy to

define protein interaction networks for peptide recognition modules. Science. **2002** 295:321-4.

- **Uetz P**, Giot L, Cagney G, Mansfie ld TA, Judson RS, Knight JR, *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. **2000** 403:623-7
- **Velankar** S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, Apweiler R, *et al.* E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res. **2005** 33:D262-5. see http://www.ebi.ac.uk/msd/index.html
- **von Mering C**, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. **2005** 33:D433-7. see http://string.embl.de/.
- **von Mering C,** Krause R, Snel B, Cornell M, Oliver SG, Fields S, *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. Nature. **2002** 417:399-403.
- **Walhout AJ**, Boulton SJ, Vidal M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. Yeast. **2000** 17:88-94
- **Wood V,** Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, *et al.* The genome sequence of Schizosaccharomyces pombe. Nature. **2002** 415:871-80.
- **Xenarios I**, Salwinski L, Duan XJ, Higney P, Kim S, Eisenberg D. DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. Nucl. Acids. Res. **2002** 30:303-305.
- **Xiao B**, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken A, *et al.* Structure of a 14-3-3 protein and implications for coordination of multiple signalling pathways. Nature. 1995 376:188-91.
- **Yaffe MB, Cantley LC.** Mapping specificity determinants for protein-protein association using protein fusions and random peptide libraries. Methods Enzymol. **2000** 328:157-70
- **Yaffe MB, Elia AE.** Phosphoserine/threonine-binding domains. Curr Opin Cell Biol. **2001** 13:131-8.
- **Yaffe MB**, Rittinger K, Volinia S, Caron PR, Aitken A, Leffers H, *et al.* The structural basis for 14-3-3:phosphopeptide binding specificity. Cell. 1997 91:961-71.

- **Yu H**, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, *et al*. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res. **2004** 14:1107-18.
- **Zanzoni A,** Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett. **2002** 513:135-40

ACKNOWLEDGMENTS

During my PhD years I benefited from the advice and collaboration of many colleagues, David Delannoy, Caius Gravilla, Gabriele Ausiello, Daniele Peluso and Elena Santonico. In particular I want to thank Andreas Zanzoni Arnaud Ceol and Maria Victoria Schneider for the friendly and fruitful collaboration in building and maintaining the MINT database.

For the achievement of this work I owe an incredible debt toward Simona Panni who carried out all the experiments mentioned and helped me writing this manuscript. I am very grateful to Federico Gherardini, Allegra Via and Manuela Helmer-Citterich who gave me great hints in handling protein patterns, to Maria Persico who helped me enormously building the protein alignments and to Luisa Castagnoli for her several valuable advices.

I am very grateful to Henning Hermjakob for his critical review of this manuscript and for the exciting collaborations we had in the last three years. I also want to thank Rolf Apweiler who hosted me at the EBI as Marie Curie fellow and gave me a lot of opportunities.

Gianni Cesareni have been since 1999 a constant source of brilliant ideas, support and encouragement. He deserves my best gratitude while MINT database and all the work I achieved before and during my PhD started on his initiative.

Finally I want to thank the very first reviewer and husband Joe Coppola for the enormous patience and kindness he had in my regards during the writing of this manuscript.