UNIVERSITÀ DEGLI STUDI DI ROMA
"TOR VERGATA"

FACOLTA' DI SCIENZE MATEMATICHE FISICHE E NATURALI

DOTTORATO DI RICERCA IN
BIOLOGIA CELLULARE E MOLECOLARE

XXI CICLO

PepspotDB: a database for the storage and analysis of experiments based on peptide array technology

Stefano Costa

A.A. 2008/2009

Docente Guida: Prof. Gianni Cesareni

Coordinatore: Prof. Gianni Cesareni

Table of Contents

1. Abstract

The mapping of the "interactome" (i.e. the network comprising all possible physical protein-protein interactions naturally occurring within a cell or an organism) of living organisms is a key asset to promote the advancement of Systems Biology. Notwithstanding the numerous insights we have gained from the study of protein-protein interaction networks, currently available interactomes present several shortcomings, one of the more crucial being the lack of information regarding the regions involved in the interactions. Especially important in this respect are several families of conserved protein domains (e. g. SH2, SH3, WW, EVH1) that mediate protein-protein interactions by binding to short linear motifs.

Our group has recently devised a strategy based on peptide array technology to study on a large scale the target recognition specificity of domains binding to short peptides. Our approach consists of an experimental and a computational part: 1) the domains are profiled by testing them with ad hoc designed peptide arrays; 2) Neural Network based predictors are trained for each of the profiled domains and the predictions are combined in a Bayesian framework with information coming from multiple orthogonal sources to obtain an integrated interaction confidence score. The approach has been applied to the identification of all human protein-protein interactions mediated by SH2 domains.

To support the projects employing our approach, we have developed a brand new database-centered application, called PepspotDB, specifically designed to facilitate the storage and analysis of molecular interaction assays exploiting peptide array technology. We hope that PepspotDB will grow enough to become a prominent resource for the storage, analysis and retrieval of peptide chip data.

PepspotDB comes with a traditional relational database, where experimental results, computational predictions and data imported from the

literature or other external sources are stored, a rich web application, providing a user-friendly, yet powerful, interface to the database, and a set of tools to automatically process raw experimental data, identify promising candidate binders and visualize sequence logos.

At the time of writing, PepspotDB contains more than 5 million records, comprising about 80 experiments and 55,548 domain-peptide interactions involving 70 SH2 domains and 7,972 unique peptides. These numbers are bound to more than double as new experiments involving other domain families are completed. Scientists studying protein-protein interactions mediated by domains recognizing linear peptides may find PepspotDB a precious resource to foster their own research.

2. Introduction

2.1. Protein-protein Interaction Networks and their Shortcomings

The completion of the sequencing of the genome of numerous organisms has helped to realize that the observed increase in genome complexity is inadequate to satisfactorily explain the greater morphological, physiological and behavioral complexity of higher organisms. Increase in the intricacy of the network that controls the expression of gene products and their interactions has been invoked as a further level of complexity that might explain the differences observed at the organismal level.

Thus, compiling an exhaustive list of the genes and the proteins they encode is merely the first step towards a full understanding of cell functioning. In the great majority of biological processes, from signal transduction to immune response, from uptake and metabolism of nutrients to cell movement, from cytoskeleton organization to regulation of gene transcription and translation, proteins and protein complexes play a key role. In carrying out their functions, proteins form either stable or transient interactions, assembling macromolecular complexes that are often essential to fulfill their biological role. Mapping all the interactions among proteins in a cell would represent a fundamental starting point to explore the mechanisms regulating living systems behavior. The network comprising all possible physical protein-protein interactions naturally occurring within a cell is called “interactome”. The interactome constitutes a scaffold to which other genomic scale data may be attached to further elucidate cell physiology.

The interactome plays a major role in stimulating the growth of a new branch of biology, called “Systems Biology”. Systems Biology aims at

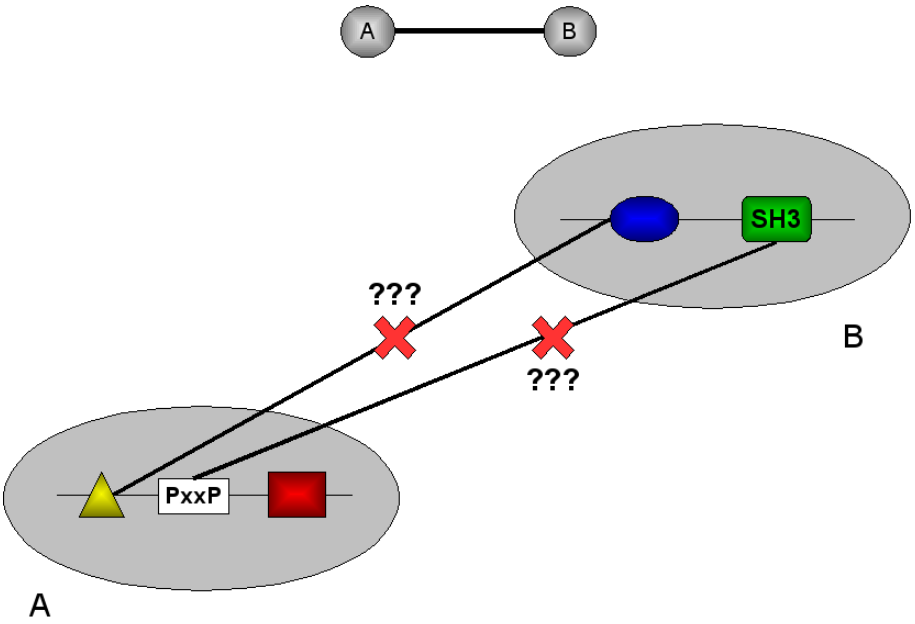


Figure 1: protein A and B are composed of three and two domains respectively. Thus, the interaction between protein A and B may be explained equally well by several possible domain-domain interactions (e.g. yellow-blue, white-green). Our ability to disrupt the A-B complex is a function of our knowledge of what domain pairs are more likely to interact.

gaining insights into the dynamics governing the behavior of complex biological systems, viewed as the result of the coordinated action of many elementary components connected to each other, eventually building an abstract model of the system capable of predicting the system's response to any variation of input (internal parameters and environmental conditions). The availability of a map of all physical connections among proteins in the

cell is mandatory to develop useful, biologically relevant mathematical models of cellular processes.

However, the information provided by standard experimental techniques for detecting protein-protein interactions may not be enough for the level of detail required by modeling. For instance, interaction data often contain no indication on post-translational modifications (PTM) that protein partners may have undergone, nor it is clear whether such modifications favor or hinder the interaction. Moreover, quantitative information on interaction affinity and kinetic parameters, crucial for describing the dynamic of assembly and disassembly processes, is only very rarely available. Finally, while the genome is essentially static and does not change during the life of an individual, the proteome and, consequently, the interactome is far more dynamic, varying its composition according to the cell cycle and environmental conditions. In other words, interaction detection techniques can only take a snapshot of the possibly occurring interactions in a particular moment of the life of a specific type of cell: by no means they provide an exhaustive description of the temporal and spatial evolution of the interactome.

Another severe limitation of current high-throughput approaches is that they are generally unsuitable for providing information on the protein regions involved in binding. Although atomic level details on protein interactions will only come from crystallographic studies of the 3D structures of multi-protein complexes (Aloy and Russell 2006), valuable insights into the mechanisms underlying the formation and functioning of protein networks can be achieved by considering the intrinsic modular nature of proteins. It is the domain composition of a protein that determines both the functions it can perform and its binding ability to form complexes with partner proteins. Elucidating the details of how proteins interact would allow us to pursue ambitious goals, such as modeling competitive binding, or designing in a rational way synthetic molecules targeted on inhibiting

specific interactions. To illustrate this concept, Figure 1 shows an interaction between two proteins, A and B, composed of three and two domains respectively. The interaction between the two proteins may be explained equally well by more than one domain-domain interaction. Supposing that we are interested in studying the functional implications of blocking the interaction between A and B, we are confronted with a question: what domains are responsible for the interaction? A wrong answer to this question may impair our ability to effectively inhibit the interaction: when a priori information about which domain pairs are more likely to interact is lacking, we have no clear indication on how to design a competitor molecule that would disrupt the A-B complex. Thus, it is clear that studying protein interactions at the domain level would provide a finer resolution description of the dynamic protein mesh in a cell and it would favor a deeper understanding of the logic underneath the wiring of protein networks.

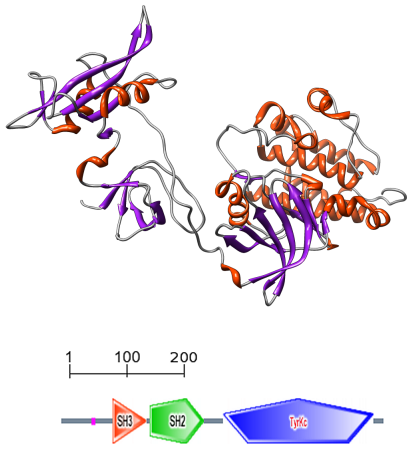


Figure 2: domain composition of the Src kinase

2.2. Proteins Can Be Decomposed into Domains

Although protein interaction graphs usually represent proteins as nodes, mathematical abstractions devoid of any structural detail, viewing a protein as a monolithic unit designed to perform a single, specific biological function is a

too simplistic description of reality. In fact, as we have mentioned, many proteins can be thought of as the assembly of smaller building blocks, termed "domains", and therefore show an intrinsic modular nature. Figure 2 shows an example of a protein composed of three clearly distinguishable domains.

Although no universally agreed definition of domain exists, protein domains are usually identified and classified according to structural, functional and evolutionary criteria. A domain can thus be described as a compact, autonomously folding structural unit, which is conserved across evolution and is capable of performing a function, independently of the context of the protein it belongs to. Proteins can be composed of one (single-domain proteins) or more (multi-domain proteins) domains. Genome-wide structural assignments of domains have shown that domain composition complexity increases in proteins of higher order organisms: about two thirds of prokaryotic proteins are multi-domain, compared to 80% of eukaryotic proteins (Liu and Rost 2004). The variety we observe in the proteomes of living organisms may have been generated from a limited set of ancestral domains with different functions. Then, the evolution of protein repertoires may have been driven primarily by three forces: 1) duplication of gene sequences coding for one or more domains; 2) sequence differentiation by mutation and genetic drift, eventually resulting in the acquisition of new functions; 3) gene recombination, which promotes rearrangements in the sequential order of domains (Chothia et al. 2003; Bauer et al. 2005). This interpretation leaves us with a question: how novel polypeptide domains have arisen in the beginning? Recently, the observation of species-specific exons in the genomes of closely related species has suggested that the "exonization" of species specific intron sequences may play an important role in the genesis of domains. Furthermore, insertions and deletions occurring in already existing protein-coding sequences, or modifications to the length of sequence repeats may contribute to the process of acquiring

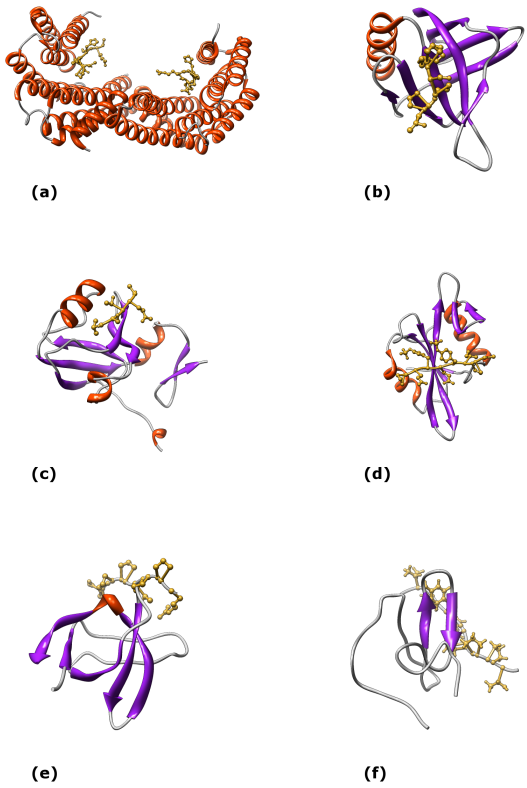


Figure 3: three-dimensional structures of six protein interaction modules. (a) 14-3-3 Zeta/Delta dimer in complex with phosphopeptides (PDB code: 1QJB). (b) EVH1 domain of Homer protein homolog 1 in complex with peptide TPPSPF derived from Metabotropic glutamate receptor 5 (PDB code: 1DDV). (c) The third PDZ domain from the synaptic protein Psd-95 in complex with a C-terminal peptide derived from CRIPT (PDB code: 1BE9). (d) Src SH2 domain bound to doubly phosphorylated peptide PQpYEpYIPI (PDB code: 1NZL). (e) C-terminal Sem-5 SH3 domain bound to a proline-rich peptide from mSos (Ac-PPPVPPRRR) (PDB code: 1SEM). (f) Yap65 (L30K mutant) WW domain in complex with GTPPPPYTVG peptide (PDB code: 1JMQ).

new structural features of increased complexity in proteins, ultimately resulting in the creation of new polypeptide domains (Schmidt and Davies 2007).

2.3. Interaction Modules: Small Globular Domains Recognizing Short Linear Motifs

Proteins associate to form either stable or transient multi-protein complexes. The assembly of stable macromolecular complexes often requires high affinity interactions, involving large portions of the protein surfaces in the formation of the contact area. Such high-energy bonds are not easily disrupted and are thus well suited to confer stability on molecular machines such as the centrosome, the proteasome and others (Brooijmans et al. 2002). However, when flexibility and responsiveness are needed, e.g. in signaling pathways, such great stability would be a limitation. In fact, in highly dynamic contexts protein-protein interactions frequently involve a modular protein domain binding to a short amino acid sequence, termed "motif" (Pawson et al. 2002).

A motif is a sequence of 3 to 12 amino acids, which acts as recognition element for a modular protein domain that binds to it. Since these motifs are thought to be unstructured in their unbound form, they are often referred to as "linear" motifs: nonetheless, they may acquire a well-defined three-dimensional structure upon binding (for example, the proline rich motifs recognized by SH3 domains adopt a left-handed poly-proline II helix conformation). Short linear motifs often lie within disordered regions of the protein, for example they may reside in exposed flexible loops.

Binding motifs are recognized by a set of conserved protein interaction modules playing a major role in many biological processes: as an example, SH3, SH2, WW, EVH1, and PDZ domains are involved in protein

trafficking and degradation, cytoskeletal organization, cell-cycle progression, cell survival and regulation of gene expression; plus, they can mediate the assembly of multi-protein complexes (PDZ domain containing proteins often act as scaffolds). Graphical representations of these and other interaction modules are illustrated in Figure 3 and their properties are summarized in Table 1. Each family of interaction modules has a defined conserved structure forming one or more 'recognition pockets', to which ligands bind. A small number of highly conserved amino acids have their side chains protruding outward in the recognition pocket. These mediate recognition of a specific signature in the ligand molecule, a short "core" motif of a few amino acids whose side chain chemical groups are complementary to those in the domain recognition pocket. Within a single domain family, the broad specificity encoded in the core motif is further refined by a few flanking residues. Thus, the residues surrounding the core motif in the target peptide contribute to define which domains it will preferentially attach to and modulate the binding affinity. Domain-motif interactions involve a small portion of the proteins' surface area and are characterized by relatively low affinities, ranging from 0.1 to 100 μM. This characteristic confers on domain-motif interactions a dynamic nature and explains their ubiquitous presence in signaling pathways, where protein complexes associate and disassociate in response to any arbitrarily varying stimulus. Some interaction modules bind their ligands dependently of some type of covalent modification, the most common being phosphorylation: for instance, SH2 domains bind with high affinity phospho-tyrosine containing peptides, while 14-3-3 proteins prefer phospho-serine containing motifs.

Domain	Size	Structure	Target Motifs	Binding Pocket	Function
14-3-3	~30 kDa proteins	9 anti-parallel helices forming an L-shaped structure	phosphothreonine or phosphoserine motifs consensus: RSxpSxP	4 helices containing hydrophobic residues form a concave amphipatic groove	regulation of many pathways (e.g. apoptosis, cell cycle)
EVH1	~110 amino acids	compact parallel beta-sandwich, closed along one edge by a long alpha-helix	proline-rich peptides consensus: E/DFPPPPXD/E	highly conserved cluster of three surface-exposed aromatic side-chains	scaffolding, signaling, nuclear transport and cytoskeletal organization
PDZ	80-90 amino acids	six beta-strands and two alpha-helices compactly arranged in a globular structure	C-terminal motifs several PDZ domains bind to phosphoinositide PIP2	elongated surface groove forms as an antiparallel beta-strand interacts with the betaB strand and the B helix	scaffolding, localization of proteins to the plasma membrane, regulation of intracellular signaling
SH2	~100 amino acids	a central hydrophobic anti-parallel beta-sheet, flanked by 2 short alpha-helices forming a compact flattened hemisphere	phosphotyrosine motifs consensus: p-YxxΨ	ligand binds perpendicular to the beta-sheet and interacts with the loop between strands 2 and 3 and a hydrophobic binding pocket that interacts with a pY+3 side chain	regulation of intracellular signalling cascades found in adaptor proteins and non-receptor tyrosine kinases
SH3	~50 amino acids	five or six beta-strands arranged as two tightly packed anti-parallel beta sheets	proline-rich motifs consensus: PxxP	flat, hydrophobic pocket consisting of three shallow grooves defined by conservative aromatic residues	signaling, cytoskeletal organization, assembly of macromolecular complexes

				ligand adopts an extended left-handed helical arrangement	found in adaptor proteins
WW	~40 amino acids	stable, triple stranded beta-sheet	phosphoserine-phosphothreonine motifs proline-rich motifs consensus: PPxY	WW or WWP name takes after the residues responsible for binding: two tryptophan residues that are spaced 20-23 amino acids apart and a conserved proline	involved in a variety of signal transduction processes

Table 1: the table summarizes the properties of six of the most relevant and well-studied interaction modules. The descriptions were taken mainly from the InterPro database, version 15.1 (Apweiler et al. 2000; Mulder et al. 2007). Ψ indicates a hydrophobic amino acid, x indicates any amino acid and p- indicates phosphorylation.

The widespread distribution of proteins containing interaction modules and the specificity of the interaction mechanism, closely resembling a key-lock model, has led scientists to postulate the existence of a 'protein recognition code', analogous to the genetic code (Sudol 1998). Such code would be composed of a set of rules, which may be encoded by relatively simple regular expressions, determining how protein interactions mediated by interaction modules can occur. However, caution should be taken in following too strictly this analogy, to avoid overlooking some relevant differences between the two. Perhaps the most noteworthy features of the genetic code are its universality and its absoluteness: a three-letter codon encodes a specific amino acid always and everywhere (or at least with very few exceptions), independently of the organism and the cellular context. Even though protein interaction modules are present in a wide range of

species, from unicellular to multi-cellular organisms, from plants to animals, their behavior is by no means universal or context-independent. The in vivo binding of a potential ligand peptide is always conditioned to local concentrations, subcellular localizations, and, in some cases, to the coordinated action of other interaction modules (e.g. adaptor proteins, Pawson 2007). Furthermore, domains belonging to the same family frequently share a considerable number of cognate ligands (Castagnoli et al. 2004), and a certain degree of overlap also exists between the target recognition rules of different families: this indicates that the protein recognition code presents a high level of degeneracy. This is known to be also an important feature of the genetic code (even if the level of redundancy in the genetic code is certainly lower), and may help confer robustness on the system by resisting to the detrimental effect of mutations.

2.4. Methods to Detect Domain-Domain Interactions

Unfortunately, no well-established experimental method is available to detect domain-domain interactions on a large scale. In principle, one could use common interaction detection techniques, e.g. yeast two hybrid (see Appendix, paragraph 6.1.1.), on collections of engineered constructs expressing only a portion of the full length protein: if a smaller construct maintains the same ability to interact as the full length protein, one may safely conclude that the region mediating the interaction lies within the domain(s) present in the fragment protein. This approach has been successfully employed by two different groups to explore the protein interaction network of two micro-organisms, P. falciparum (Lacount et al. 2005) and H. pylori (Rain et al. 2001), by two hybrid screenings of a library of protein fragments.

However, generally speaking, determining the domains mediating a

protein interaction is a time consuming task. This prompted the development of several computational methods to identify pairs of putative interacting domains. Although many different algorithms have been devised for this purpose, so far most of them rely on the same basic assumption: if a pair of domains co-occur in interacting protein pairs significantly more frequently than in non-interacting ones, they are likely to interact (see Figure 4). Based on this hypothesis, statistical methods may be employed to search domain pairs frequently recurring in interacting protein pairs.

Sprinzak and Margalit (Sprinzak and Margalit 2001) scored putative interacting domain pairs by computing the log-odds of the two domains co-occurring in interacting pairs to the co-occurrence expected on a random base. Ng et al. (2003) developed a scoring system aimed at integrating the information from protein-protein interactions, multi-protein complexes and domain fusion events. The results of their predictions were stored in an online database called InterDom (http://interdom.lit.org.sg). Nye et al. (2005) adopted a rigorous statistical approach and applied a sophisticated simulation technique to assign to each pair of domain superfamilies occurring in a generic protein interaction dataset a p-value reflecting the likelihood that they are able to interact. Deng et al. (2002) developed a Maximum Likelihood Estimation (MLE) and an Expectation Maximization (EM) algorithm to infer probabilities of the domain interactions underlying a set of protein interactions; in a more recent paper (Lee et al. 2006), they extended their method by integrating interaction probabilities with information from protein fusions and Gene Ontology (Ashburner et al. 2000) functions through a Bayesian approach. Riley et al. (2005) modified the first version of the algorithm by Deng et al. (2002) and improved it by introducing the E-score, a measure reflecting the importance of a specific domain-domain interaction to explain a set of protein-protein interactions. Jothi et al. (2006) opted for a different approach, looking at the relative degree of co-evolution of domains in interacting protein pairs: they provided

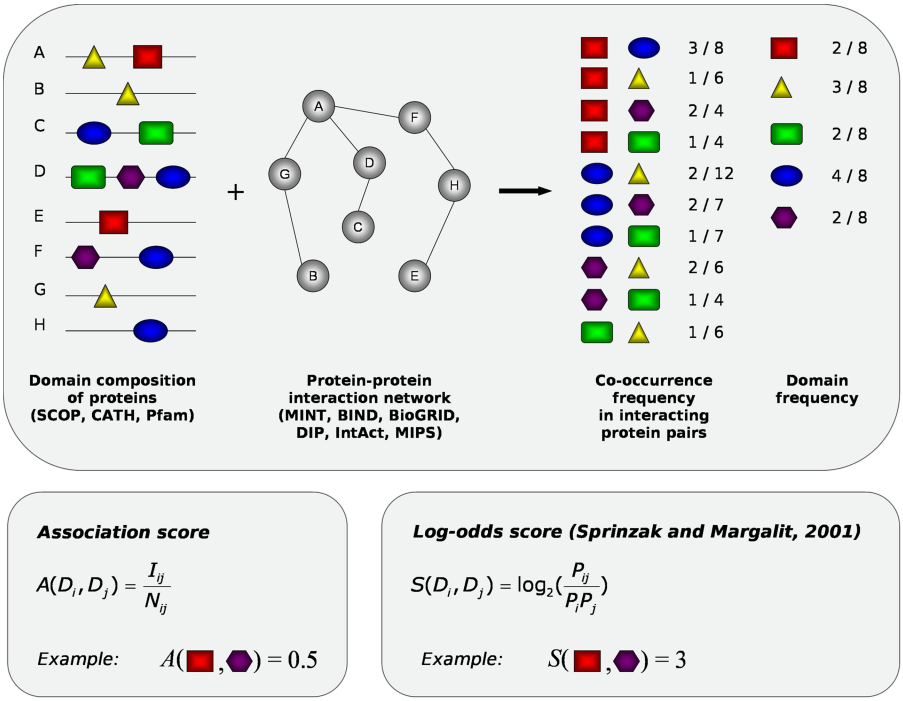


Figure 4: The figure shows the basic idea underlying algorithms to infer domain-domain interactions. The algorithms need as input a protein interaction network along with information about the domain composition of the proteins present in the network. Given these data, it is possible to compute the co-occurrence frequency of domain pairs in interacting protein pairs as the ratio between the number of interacting protein pairs containing the domain pair of interest and the total number of protein pairs containing it. The domain frequency in the set of proteins appearing in the protein network can be also calculated. Next, these frequencies can be combined to assign domain pairs an interaction probability. The lower part of the picture shows two different methods to rank domain pairs according to the likelihood that they interact: the fairly simple association score and the log-ratio score devised by Sprinzak and Margalit. The first method simply ranks domain pairs by their co-occurrence frequency in interacting protein pairs, whereas the second relates this value to the co-occurrence frequency that would be expected on a random base (P_i represents the frequency of domain i in the proteome). Several algorithms of various complexity levels have extended this basic procedure to improve the accuracy and reliability of the predicted domain-domain interactions.

evidence that pairs of domains mediating the protein interaction are more likely to co-evolve with respect to non-interacting domain pairs.

Although some of the aforementioned methods show promising results, all of them are far from perfect. Since all methods invariantly require as input a protein interaction dataset, their performance is strictly dependent on the quality of the interaction data, which are often affected by high false positive and false negative rates. Another major issue is validation: how the reliability of predictions can be assessed? Accurate estimation of the algorithm's performance would recommend the comparison of the output predictions with a reference set of trusted positives and negatives. The definition of such reference set is greatly impaired by the scarcity of known interacting and non-interacting domain pairs. It is common practice to consider true interacting domain pairs those which have been observed to interact at least in one solved three-dimensional structure and true non-interacting pairs those which never come into contact in known structures. Given the small number of non-redundant three-dimensional structure contained in the PDB (Protein Data Bank, Berman et al. 2000), many true interacting domain pairs may not be represented at all in the positive set, whereas the number of non-interacting pairs in the negative set is likely to be overestimated. Accuracy of the reference set could also be questioned: in many cases it is hard to tell whether two residues come close because a biologically meaningful interaction between the two has occurred or merely due to crystal packing. For this reason, Shoemaker et al. have recently developed a strategy, based on structural criteria, to discriminate biologically relevant protein domain interactions from artifactual ones (Shoemaker et al. 2005).

Underneath these computational approaches lies the assumption that protein interactions can always be explained postulating that one of the globular domains composing the first interaction partner binds to one of the globular domains composing the second interaction partner. Unfortunately,

there are instances in which this assumption does not hold: in some cases, the interaction between the protein partners may involve more than one globular domain per protein; furthermore, when the interaction is mediated by a domain recognizing short linear motifs, which often reside in unstructured, disordered regions, outside of globular domains, no globular domain of the linear motif containing protein may be involved in the interaction at all. This may well be the reason why well-known motif binding domains, like SH2, SH3, WW, GYF, EVH1, are sometimes inferred to bind to structural domains that do not contain the corresponding target motif. Indeed, it is difficult to improve the existing computational algorithms for domain-domain interaction prediction by taking into account domain-linear motif co-occurrences in interacting protein pairs, because linear motif matches are very abundant in the proteome of any organism and the vast majority of their occurrences, being due to chance, is likely irrelevant from a functional perspective and deleterious to the effectiveness of the statistical reasoning of the algorithm. Thus, purely experimental approaches or mixed computational/experimental approaches, if supported by solid experimental data, are advisable for the study of domain-peptide interactions.

2.4.1. Methods to Detect Domain-Peptide Interactions

In trying to unveil the rules governing the mechanisms of interaction between protein interaction modules and peptides, one would like to explore the sequence space as exhaustively as possible, ideally testing the binding of all representatives of each domain family occurring in the proteome of a living cell to all possible amino acid sequences of reasonable length.

However, this brute force strategy is not feasible due to the technical limitations of current techniques. Fortunately, the impossibility to reach perfect generality and completeness in searching potential ligands may be

partially overcome by taking into account some a priori knowledge about the binding determinants of any given domain.

Capitalizing on the characterization of a large number of target peptides and on the detailed structural information contained in high-resolution three-dimensional structures of interaction modules in complex with their ligand peptides, the sequence characteristics of binding peptides can be determined and confirmed by mutagenic analysis.

This information defines a biased sequence space where a few pre-determined positions contain specific amino acids while the others are allowed to vary in a combinatorial fashion. Thus, the new search space is reduced in size and can be explored experimentally with currently available techniques. This approach may be concretized by constructing an "oriented peptide library", a biased collection of peptides of degenerate sequence but having fixed amino acids in the "orienting" positions (e.g. for SH2 domains, the phospho-tyrosine residue would be the orienting amino acid, whereas for SH3 domains the two proline in the PxxP motif) (Yaffe and Cantley 2000). The peptide mixture is then incubated with the domain of interest. Subsequent sequencing of the adsorbed peptides allows to determine the positions showing enrichment for any particular amino acid. Although very powerful in theory, this approach has not been widely employed, mainly because of the technical expertise required to perform complicated peptide biochemistry (Santonico et al. 2005).

An alternative approach is based on phage display (Scott and Smith 1990): a library containing a large number (in the range of 10^9-10^{10}) of short (10-15 amino acids) random peptides is displayed on bacteriophage capsids and is panned against a domain used as bait. The peptides having affinity with the domain can be purified and their sequence can be easily obtained by sequencing the DNA of the capsid gene. It is then possible to derive a consensus sequence by aligning a reasonable number of interacting peptides. Although this technique has been successfully employed to profile several

families of interaction modules (Sparks et al. 1994; Rickles et al. 1994; Vaccaro et al. 2001; Mongiovì et al. 1999; Cestra et al. 1999; Paoluzi et al. 1998; Dente et al. 1997; Freund et al. 2003), there are some limitations to it: first, whereas it is fairly easy to identify which residues are highly conserved among ligands, it is often problematic to uncover statistical correlations in less conserved positions; second, some peptides are actually able to bind to the domain, even if they do not perfectly match the consensus. When the derived consensus is used to infer physiological partners in the proteome, these two problems inevitably give rise to the occurrence of false positives and false negatives.

A third approach, called SPOT synthesis (Frank 2002), successfully addresses the limitations of phage display. The technique is based on the chemical synthesis of a high number of peptides on a cellulose membrane or a glass slide, in an array format: the binding of the domain to each spot is detected by a fluorescent probe, whose intensity is measured by a laser scanner. Since binding of each and every peptide in the collection is tested independently and is semi-quantitatively described by a figure correlating with the dissociation constant, we are able to tell readily which peptides bind and which do not, eliminating the inference step that represents the major drawback of the previously described techniques.

However, the number of peptides that can be spotted together on a chip with current facilities is in the order of 10^4, limiting the applicability of this approach to a rather restricted search space. Recently, the complementarity of the phage-display and the SPOT synthesis approaches has been exploited in the design of a two-step strategy, named WISE (Whole Interactome Scanning Experiment) (Landgraf et al. 2004), that aims at combining the strengths of the two methods into a single general purpose procedure (see Figure 5). The first step involves the definition of a “strict” consensus sequence by panning a phage displayed peptide library against the selected domain. Then, a “relaxed” consensus is obtained with the aid of

computational tools and it is used to select a number of peptides between 5,000 and 10,000 from all the protein sequences in the proteome that match the consensus.

Finally, the peptides are synthesized on a chip and are probed with the domain of interest. The selection step makes it possible to test all potential domain targets in a proteome.

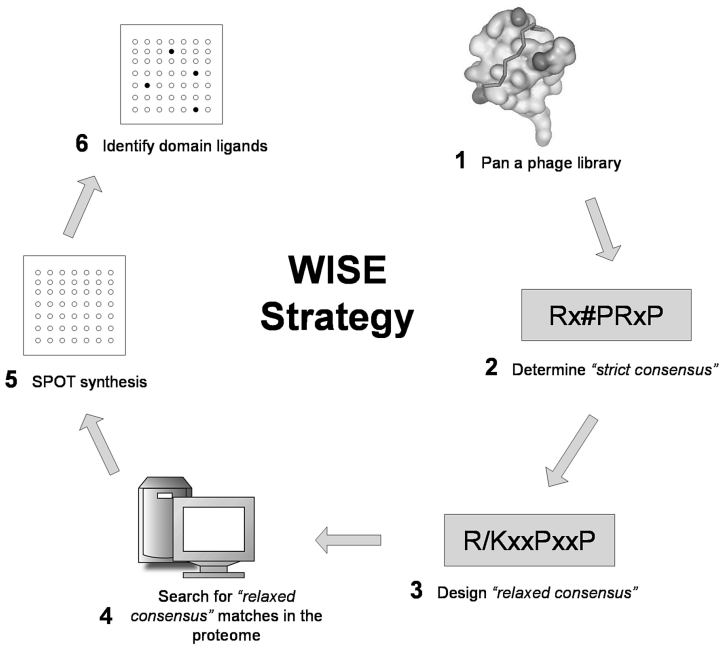


Figure 5: overview of the WISE strategy.

2.5. Public Databases Containing Domain-Domain and Domain-Peptide Interactions

Given the paucity of available experimental information about domain-domain interactions, it is not surprising that only a few repositories of domain-domain interactions exist, and those few that are there are either limited in scope or heavily dependent on computational predictions, in order to achieve greater coverage. iPfam (Finn et al. 2005) falls within the first category, since it is limited to the accurate description (at the resolution of residue-residue contacts) of domain–domain interactions observed in protein complexes whose 3D structure is known. Examples of the second category are InterDom (Ng et al. 2003b) and DIMA (Pagel et al. 2006), two databases that, albeit differing in scope, pursue the same objective: the assembly of a global domain interaction graph linking domains that are likely to interact, through the integration of multiple data sources and prediction techniques. DOMINE (Raghavachari et al. 2008), a database of known and predicted domain-domain interactions inferred from PDB entries and by 8 different computational approaches, operates along the same lines.

The situation is even worse when we narrow down to the study of domain-peptide interactions, that is, when we are interested in discovering the physiological targets of protein domains that recognize short linear motifs. To this day, the only large scale effort to catalog all experimentally observed domain-peptide interactions has been undertaken by our group with the development of DOMINO, a database of protein interactions mediated by motif-binding domains (Ceol et al. 2007). Other attempts towards this direction have been made, but they are more focused on certain classes of domains, such as PepCyber (Gong et al. 2008), a database of human protein-protein interactions mediated by phosphoprotein binding domains (PPBDs), or PDZBase (Beuming et al. 2005), a manually curated protein-protein interaction database developed specifically for interactions involving PDZ

domains. Several groups have tackled the problem from a different perspective and, instead of building a database where domain-peptide interaction records would be stored, have designed prediction algorithms that would estimate the likelihood of any query sequence to be targeted by a specific domain. SMALI (Li et al. 2008), iSpot (Brannetti et al. 2001) and ScanSite (Obenauer et al. 2003) are examples of this approach, that has the advantage of giving an indication on the likelihood of a candidate interaction to occur even when no experimental information is available.

3. Project Description

3.1. The Context: Three Projects Studying the Specificty of Classes of Domains Recognizing Short Linear Motifs and a Class of Enzymes Recognizing Phosphorylated Substrates

Our group has been studying protein interactions mediated by domains that bind to short linear motifs for several years. We have embarked on three big projects, respectively aiming at: 1) the mapping of all the interactions between human proteins mediated by SH2 domains; 2) the mapping of all the interactions between human proteins mediated by domains preferentially binding to proline-rich motifs (e.g. SH3, WW, GYF and EVH1 domain families); 3) the characterization of the substrate recognition specificity of a representative number of human protein tyrosine phosphatases (PTPs). Unlike the other projects, the third project is not focused on discovering physical interactions between proteins but rather on identifying novel enzymatic targets of tyrosine phosphatases. However, as will be clarified later on, the adopted experimental approach, based on peptide array technology (see Appendix, paragraph 6.1.3.), is essentially the same as the other projects.

Our group, with the contribution of our partner Jerini AG, has devised a strategy based on peptide array technology to study domain-peptide interactions on a large scale: I) a membrane is prepared before peptide synthesis and treated to ensure the covalent binding of the C-terminus of each peptide; II) the peptides are automatically synthesized on the cellulose membrane using a SPOT synthesizer under positional control of the LISA software (Jerini AG, Berlin, Germany); III) the freshly synthesized peptides

are transferred on a microscope glass slide in a high density array format, composed of up to 20,000 spots; IV) the glass slide is incubated with the domain whose target recognition specificity is being investigated, expressed as a GST-fusion protein; V) the slide is washed and the intensity of the signal emitted by each spot is measured by a laser scanner to detect successful binding reactions. The technology imposes a fairly strict constraint on the length of the peptides to be spotted on the chip, but, considering the particular domain families that we intend to study, namely those domains that recognize short linear motifs (within the range of 9-15 amino acids), we may safely conclude that our experimental approach is adequate for the given purpose.

We have recently extended this approach to make it suitable to the study of the substrate recognition specificity of human protein tyrosine phosphatases (PTPs). The characterization of the phosphatase/target network is difficult because of the transient nature of the enzymatic interaction between phosphatases and their substrates. To overcome this limitation, a mutant form of this enzyme class, dubbed trapping mutant, was developed. This mutant maintains the wild-type substrate specificity while reducing the catalytic activity (Blanchetot et al. 2005). Lacking the ability to carry the dephosphorylation reaction through, the mutant remains attached to its substrate long enough for us to detect the interaction.

Tyrosine phosphatases, by definition, carry out their enzymatic activity on phosphorylated tyrosine residue. It is also well-established that all SH2 domains, albeit they differ somewhat in specificity among one another, recognize and preferentially bind to target sequences containing a phosphorylated tyrosine residue. Thus, to identify the targets of SH2 domains and tyrosine phosphatases on a proteome-wide scale, we needed an array containing a representative sample of the human phosphoproteome. In order to compile the list of peptides to be synthesized, we scanned the literature and publicly available databases looking for experimentally

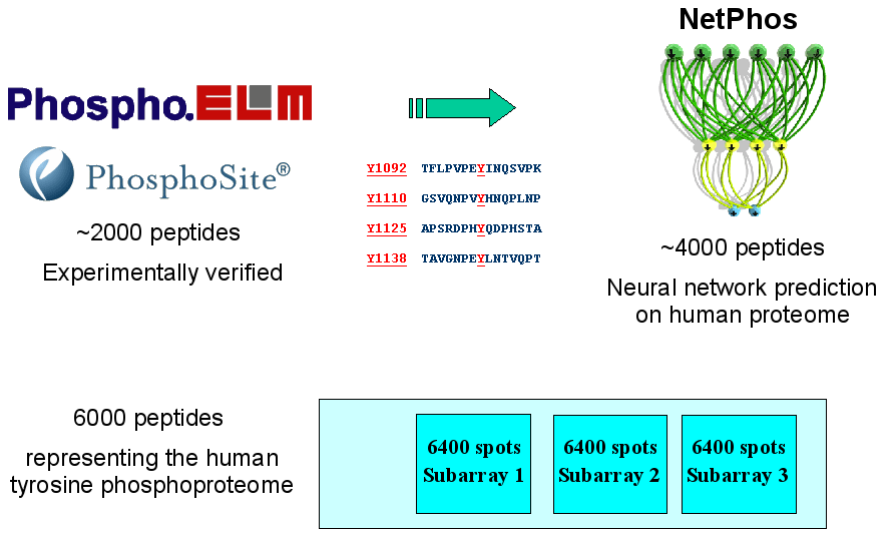


Figure 6: overview of the steps undertaken to define the array of peptides containing a phosphorylated tyrosine residue that would be used to study the interaction partner recognition specificity of SH2 domains and the substrate recognition specificity of Protein Tyrosine Phosphatases: 1) ~2,000 experimentally observed tyrosine phosphorylation sites were extracted from the literature and publicly available databases; 2) NetPhos (Blom et al. 1999), a neural network based predictor of phosphorylation sites, was run on the entire human proteome and produced an additional list of ~4,000 predicted tyrosine phosphorylation sites. The analysis yielded a total of approximately 6,000 peptides, that were spotted in triplicate on the chip.

observed tyrosine phosphorylation sites. On top of that, we ran NetPhos (Blom et al. 1999), a neural network based predictor of phosphorylation sites, on the entire human proteome. The analysis produced a total of approximately 6,000 peptides, that were spotted in triplicate on the chip. This is the reference array for projects 1 and 3 (Figure 6). Although the array

containing phosphotyrosine peptides was initially designed to cover most of the phosphoproteome known when this project started, the recent development of mass spectrometry based technology for the identification of phosphopeptides has stirred an explosion of novel information and the collection of phosphorylated peptides contained in databases (Diella et al. 2008; Hornbeck et al. 2004) now largely exceeds the number of peptides represented in our array. Thus, in order to be able to offer a resource that can reliably infer the SH2 ligands of any recently discovered peptide, we decided to develop Artificial Neural Network (ANN) predictors for each of the SH2 domains. Predictors for each of the tyrosine phosphatases are currently in an early stage of development.

A different array was designed for the study of domains recognizing proline-rich motifs. Since proline containing sequences are very abundant in the human proteome and their number greatly exceeds the maximum number of peptides that can be spotted on a glass slide with the currently available technology, we introduced a target selection step prior to synthesis, following the WISE approach described earlier. This allowed us to shrink the array size to approximately 9,000 sequences, that were subsequently spotted in duplicate on the chip.

From the experimental observation that some domain has the biochemical potential to bind to a particular peptide spotted on the chip, we may feel safe drawing the conclusion that the proteins containing respectively the domain and the peptide are capable of interacting in vivo. However, this inference may 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 it could be inaccessible to the 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. To assess the reliability of a candidate domain peptide interaction in its in vivo context, albeit only indirectly through circumstantial evidence, and to complement the *in vitro* evidence coming from chip data analysis with other pieces of evidence gathered from multiple independent sources, we have developed a Bayesian framework that integrates our predictions with information about protein co-localization and co-expression, level of disorder in the protein containing the peptide, degree of conservation of the binding site in related species, and distance between the supposedly interacting proteins in the human interactome. Thanks to the Bayesian framework, all these orthogonal information sources could be seamlessly combined to obtain a single global interaction confidence score.

If we consider that the outcome of the profiling of just one domain includes thousands of signal intensity measures, binding predictions and global interaction confidence scores, the latter obtained integrating in a Bayesian framework experimental results, computational predictions and contextual evidence, it is evident that each of the three proteome wide projects that have been just described is going to produce an enormous amount of data. Optimized storage, careful analysis and quick and simple interrogation of such an extended dataset is therefore unconceivable without the aid of a properly designed data management system and a set of computational tools developed ad-hoc.

3.2. PepspotDB: Aim of the Project

The pressing need to find a practical solution to the data management problem outlined in the previous paragraph convinced us to start a new project, which is the subject of this thesis.

The main purpose of this project is to develop a new database, called

PepspotDB, specifically designed to store in a single integrated resource the outcome of the three big projects described earlier, and, more generally, of all experiments concerning molecular interaction assays exploiting molecular array technology, along with publicly available information relevant to the questions we set out to investigate, in order to facilitate the fruition of the data and the formulation of new testable biological hypotheses from them. We decided to develop a brand new database, because we deemed readily available solutions, albeit certainly fitting to their own purpose, either too broad (MINT (Zanzoni et al. 2002), DOMINO (Ceol et al. 2007), both developed by our group), or too narrow (PepCyber, PDZBase) in scope. Although PepspotDB has been developed primarily to support research activities within our group, we plan to open it to the community, hoping that within a few years it will grow enough to become a key reference for the storage and retrieval of peptide chip data.

PepspotDB contains five main data types:

(a) Experimental measures
(b) Neural Network predictions
(c) Binary interactions
(d) Bayesian scores
(e) Phosphorylation sites

Among all domain-peptide pairs that are tested either experimentally or computationally, only those with signal intensity or prediction score above a certain threshold are considered strong evidence of interaction. These candidate binary interactions are singled out and stored in the database in a separate table for easy visualization and retrieval.

The Bayesian integration of all orthogonal pieces of contextual evidence with the predictions obtained from experimental data provides a Bayesian score for each tested domain-peptide pair.

Since both SH2 domains and phosphatases act upon peptides containing phospho-tyrosine residues, and since other families of interaction domains

also recognize phosphorylated residues, we decided to incorporate into the database information on experimentally determined phosphorylation sites, that were imported from the two most important publicly available databases hosting phosphorylation data, PhosphoSite (Hornbeck et al. 2004) and Phospho.ELM (Diella et al. 2004). Information on protein sequences was also imported from an external database, UniProtKB (The UniProt Consortium 2007, The UniProt Consortium 2009), the reference repository for protein records.

In the next paragraph, we will describe briefly, with the help of UML class diagrams, PepspotDB's data model, illustrating the key design principles and practical exigencies that have driven its development.

3.3. Database Design

3.3.1. UML class diagrams

Fig. 7-10 resort to the formalism defined by the UML Specification (http://www.omg.org/spec/UML/2.1.2) for class diagrams to visualize relevant portions of the more extended PepspotDB model. In the Unified Modeling Language (UML), a class diagram is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, and the relationships between the classes. In computer programming languages following the Object Oriented Programming (OOP) pattern, a class represents the abstraction of a concept; objects are particular instances of a certain class, that is, distinct actualizations of the one concept represented by the class. Thus, a class in the diagram purports the corresponding idea or concept existing in the particular domain of the real world we are modeling. For example, the

"Protein" class stands for real life proteins in the sense that it symbolizes the essence of a protein. Borrowing the terminology from computer science, it can be said that each particular protein is an object (or instance) of the "Protein" class.

A class is composed of a state plus the operations manipulating its state. If the state of a class is made permanent by coupling it with a record in a database, the class can be called entity. A class can be said to "inherit from" another class. What is meant by this expression is that the "heir" class, as it were, shares all the features of its ancestor (attributes and operations) and adds some more that are its own only: it is a sort of more specialized version of the same concept. Relationships between classes are called associations.

As to the graphical representation of UML class diagrams, a class is drawn simply as a rectangle divided in three compartments: name, attributes and operations (the third compartment is omitted in Figures 7-10 for the sake of clarity). Different graphic symbols are employed to distinguish among different kinds of relationship, such as mere association, composition (black diamond), aggregation (white diamond) or inheritance (triangular arrowhead), but they all appear as solid lines connecting two classes.

This brief description should suffice to understand the class diagrams shown in Fig. 7-10 and the descriptions thereof.

3.3.2. Requirements and Guidelines

The most important requirement that we were anxious to meet was flexibility, which was key to assure the database would be able to grow smoothly, with no need for major redesign steps, as the number and diversity of projects receiving support from it increased. On the other hand, an all-embracing database would have been inevitably poorly optimized for any particular application and thus most likely prone to performance issues.

The integration of external data sources was another, yet related, capital issue. The two most common alternative integration strategies are “deep” and “shallow” integration. Deep integration basically consists in replicating the information stored in an external data source into the in-house database; it requires the translation or mapping of the external database's data model to the data model of the destination database and periodic updates to bring the local copy of the data in line with the original source. Shallow integration is much easier to accomplish: the data remain in the primary data source (no replication occurs) and some sort of hyperlink pointing to them is created in the secondary data source. The advantage of the first solution lies in the much superior performance achievable in terms of response times: in fact, since the data are mirrored locally, no additional network traffic is generated. Moreover, the external data are nicely fitted into a unifying schema with respect to which they no longer appear as external, guaranteeing tight and smooth integration. On the other hand, deep integration presupposes a considerable design effort to work out a more complex unified schema and comes with major maintenance issues. Shallow integration imposes a much lighter burden on the database engineer's shoulders, often involving nothing more than the storage of a Unified Resource Locator (URL) in the proper table column. However, this second solution presents serious drawbacks in terms of performance, due to network latency and possible failure, and also of flexibility, because external data can not be directly manipulated in this approach, but are rather simply referred to. Maintenance issues are also not completely absent (URLs should be updated once in a while lest broken links may occur).

Our design choices were oriented towards the achievement of the most suitable trade-off between flexibility and performance. We renounced the idea of developing yet another general purpose molecular interaction database, since there were plenty of good ones around (e.g. IntAct (Hermjakob et al. 2004a), DIP (Xenarios et al. 2000), MINT, DOMINO) and

our focus was more pointed towards experiments based on peptide array technology. Thus, we thought that developing a database that could be used to store and analyze molecular interaction assays based on any kind of molecular array (not just peptide array) would be a satisfactory solution, with a scope neither too narrow nor too wide. As for the integration of external data sources, we did not fully commit ourselves to either strategy, but we decided on a case by case basis which one was most appropriate to gain performance without losing too much flexibility. Hence, for instance, we chose deep integration to import protein data from UniProtKB (The UniProt Consortium 2007; The UniProt Consortium 2009) and phosphorylation data from Phospho.ELM (Diella et al. 2004) and PhosphoSite (Hornbeck et al. 2004), whereas shallow integration seemed to us more effective for linking binary interactions stored in PepspotDB with relevant records in our general purpose Molecular INTeraction database (MINT, Zanzoni et al. 2002).

3.3.3. Modeling Entities

The entities composing PepspotDB's data model are organized according to one fundamental hierarchy (see Fig. 7). At the top of the hierarchy we find the DbRecordObject entity, representing any object stored in the database. Each record stored in the database is associated to a database User, holding ownership upon it. This allows to define access privileges at the record level: for example, an application built on top of PepspotDB might choose to block reading or writing from users that do not own the record.

Moving down the hierarchy one more level, we encounter the concept of AnnotatedObject. AnnotatedObjects differ from DbRecordObjects in that they can be associated to Annotations and cross-references (Xref class). The Xref class acts as a bridge between PepspotDB and foreign DataSources:

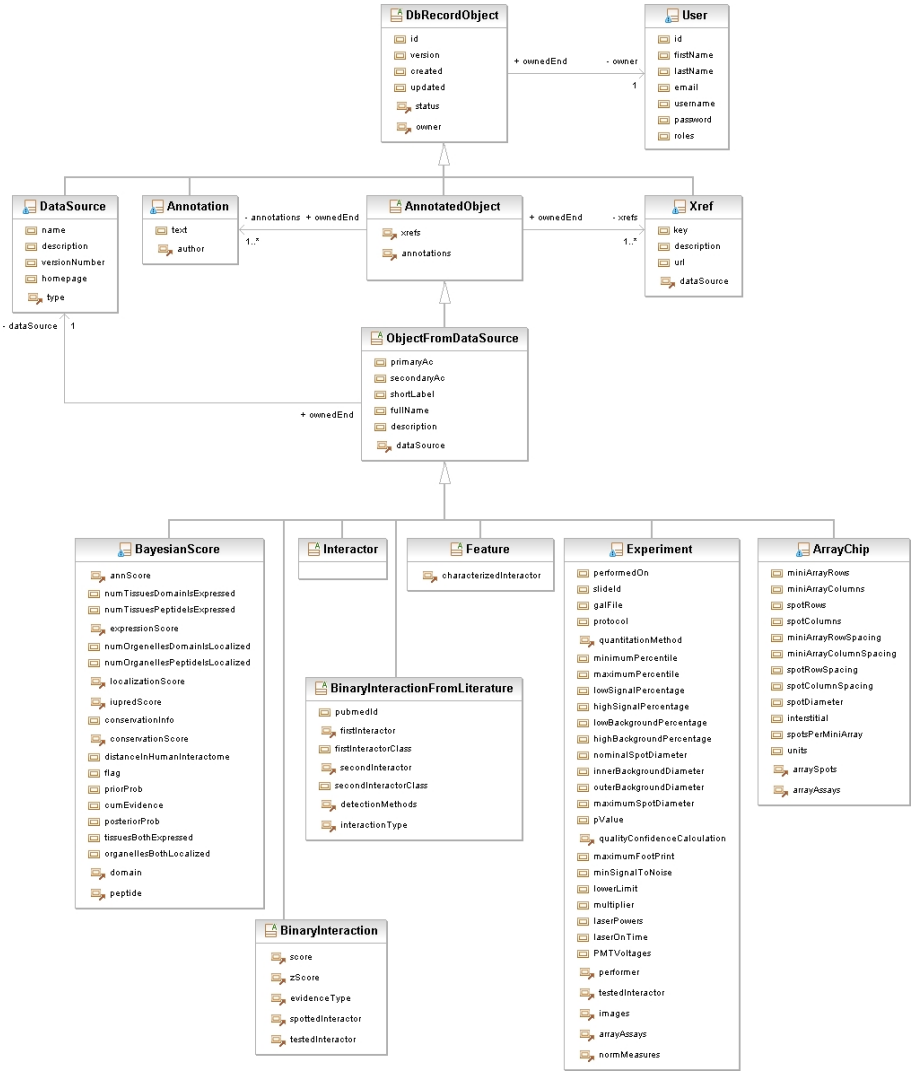


Figure 7: UML class diagram showing the one fundamental hierarchy according to which entities in PepspotDB's data model are organized.

external records can be straightforwardly referenced (shallow integration) simply by wrapping their Universal Resource Locator (URL) in a Xref object, through which they are associated to an instance of AnnotatedObject existing in PepspotDB. This design guarantees maximum flexibility, since almost any object in the model derives from AnnotatedObject and thus can be associated to cross-references pointing to external records. As an application example, shallow integration was instrumental to integrating records from PhosphoSite (Hornbeck et al. 2004) and Phospho.ELM (Diella et al. 2004).

However, deep integration was also employed: in fact, the next specialization level down the hierarchy of entities is occupied by the ObjectFromDataSource entity, which epitomizes a record, of whatever nature, coming from a DataSource (e.g. an external database or PepspotDB itself). A special DataSource describing PepspotDB itself was created and linked to all objects that, being originally created in PepspotDB, were not imported from any external DataSource. ObjectFromDataSource may be regarded as a junction point in the data model: specialized entities can be derived from it and can be used to integrate into the schema entities that were imported from separate DataSources. For example, protein records were imported from UniProtKB and transformed in Protein entities. Protein inherits from Interactor and its attributes are tailored to accommodate relevant pieces of information contained in the original record. The same was done with Post-Translational Modification (PTM, inherits from Feature) entities: PhosphoSite and Phospho.ELM records could be easily mapped to attributes of the PTM class. The last example also illustrates that nothing forbids external data sources be integrated with both deep and shallow strategies. Instances of ObjectFromDataSource and its descendant classes are all characterized by a primary accession number (primaryAc), identifying the record in the DataSource it comes from, and possibly multiple secondary accession numbers.

DbRecordObject, AnnotatedObject and ObjectFromDataSource form the backbone of the data model. From these, all other entities branch out and differentiate further: DataSource, Annotation, Xref, BayesianScore, Interactor, BinaryInteraction, BinaryInteractionFromLiterature, Feature, ArrayChip, Experiment. In the rest of the paragraph, we will zoom in on crucial areas of the data model to explain in greater detail how experiments and interactions are modeled in PepspotDB.

3.3.4. Modeling Experiments

Although the database was originally conceived as a repository of experiments employing peptide arrays to detect domain-peptide interactions, we thought it expedient to make the database structure as general as possible, capable of accommodating virtually any kind of interaction assay based on array technology.

Figure 8 shows a class diagram portraying a subset of classes, extracted from the larger PepspotDB data model, capturing the concept of experiment and the concepts related to it. First of all, it should be noted that PepspotDB assumes that any experiment involves the probing of one or more arrays of some kind with a potential interactor. In other words, PepspotDB assumes experiments are based on array technology. This, of course, imposes a strict constraint to the number of different types of experiments that can be stored in the database, but, on the other hand, it opens the way for a database structure specifically tailored for these kinds of experiment, thus allowing PepspotDB to organize the data more effectively and with a greater level of detail than other more general purpose databases of molecular interactions, such as MINT. More specifically, an Experiment is composed of one or more ArrayAssays, each one referring to a specific ArrayChip and identifying a group of Measures, each of which is related to a single

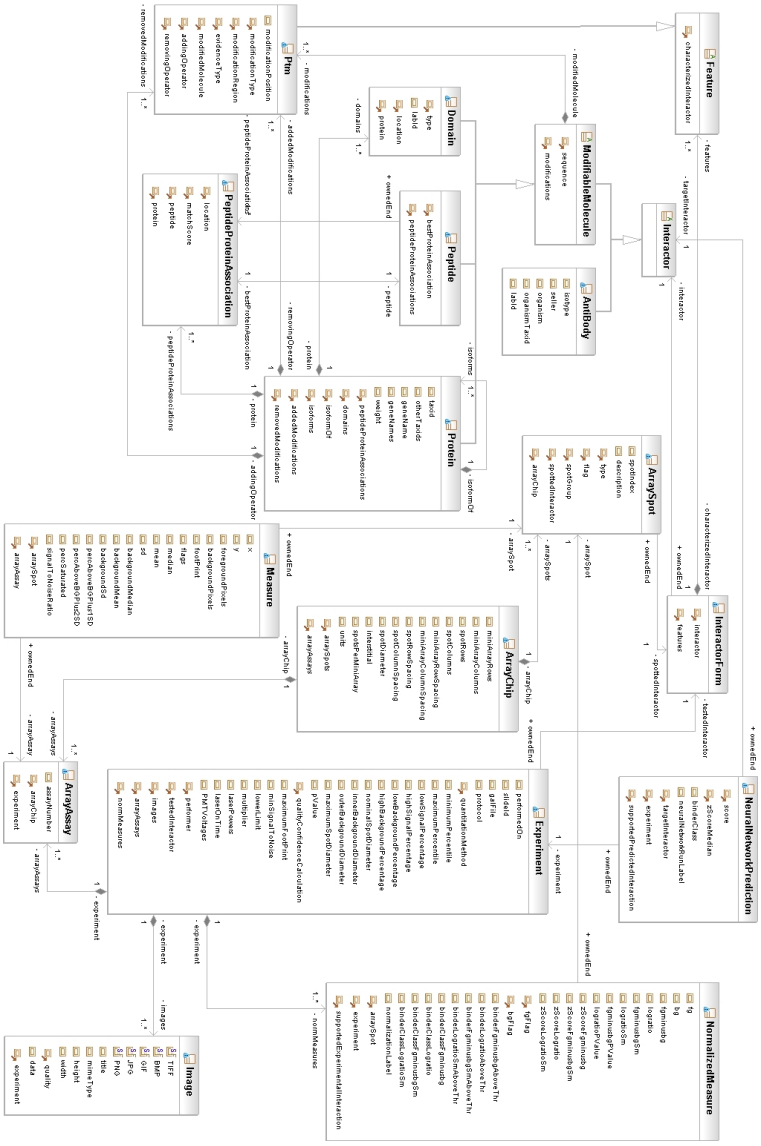


Figure 8: UML class diagram portraying the entities that contribute to the modeling of an experiment.

ArraySpot (see Fig. 8). ArrayChips are composed of ArraySpots and represent a particular array layout. ArraySpots correspond to positions in the array and can be either control spots, marker spots (used by the scanner software to align the quantization grid), blank spots or spots containing an Interactor. A Measure object encapsulates most of the data produced by the scanner software upon quantization of a particular array spot.

Experiments are often performed in duplicate or in triplicate: different replicates of the same array layout within the same experiment are modeled by different ArrayAssay objects, one for each replicate, referring to the same ArrayChip and Experiment. Thus, in an experiment performed in triplicate, three Measures for each ArraySpot will be associated to the respective ArrayAssay object representing one replicate of the same ArrayChip.

The Interactor class conveys the general idea of a molecule that can interact with some other molecule, abstracting from the actual type of molecule we are talking about (e.g. a protein, a region of a protein, a nucleic acid, etc...). ModifiableMolecule is a particular class of Interactor provided with a sequence to which one or more PTM (Post-Translational Modification) objects can be attached. ModifiableMolecule is further specialized into Peptide, Domain and Protein. A Peptide can be associated to many proteins through PeptideProteinAssociation entities (most often a peptide matches the sequence of several homologous proteins); a Protein in turn can be composed of many domains. These layers of abstraction allow PepspotDB to deal with experiments of different nature exactly in the same manner: since experiments are modeled as interaction assays testing, in one-versus-all fashion, a generic Interactor against one or more arrays composed of ArraySpots containing also generic Interactors, domain-peptide, domain-domain, antibody-antigen or protein-protein interaction assays are viewed as perfectly equal.

Furthermore, PepspotDB foresees the possibility that an Interactor may participate in different experiments in modified forms, e.g. it could be

mutated in one experiment and phosphorylated in another: InteractorForm objects signify the particular configurations assumed by Interactors in the context of a specific experiment and are used to create experiment-wise variants of an Interactor by associating to it one or more Features (the only implemented one so far being PTM, Post-Translational Modification). This is the reason why the ArraySpot class is associated to the Interactor class through the InteractorForm class and not immediately. The same is true for the Experiment class. To illustrate this concept with an example: let's suppose we wanted to capture the fact that protein A participates in experiment X in its canonical form and in experiment Y in a phosphorylated form and no InteractorForm class was present; we would have to create two distinct Protein objects, one representing A in its canonical form and another representing A in its phosphorylated form and link them to the respective experiments, thus unnecessarily duplicating information. With the introduction of the InteractorForm class, we can now create two InteractorForms, associate one of them to the proper post-translational modification (PTM), and link both forms to the same Protein object corresponding to protein A.

Besides raw data, i.e. the figures output by the laser scanner, encapsulated by the Measure object, PepspotDB's data model allows the storage also of experimental data after processing, a procedure during which filters are applied to the data to attenuate noise and the redundant information contained in the two or more replicates is collapsed into one figure. In a typical triplicated experiment, the three Measures associated to a particular spot undergo processing and are merged into a single NormalizedMeasure instance relative to the spot. As was mentioned earlier, PepspotDB encompasses not only experimental observations, but also computational predictions obtained from Neural Network predictors. The NeuralNetworkPrediction class models such computational predictions, linking the target Interactor for which the prediction was made to the

Experiment whose data were used to train the Neural Network predictor. Finally, scanned images (Image class) of the chip may be stored in the database and linked to their respective experiment.

3.3.5. Modeling Interactions

As was explained earlier, an experiment based on molecular array technology consists in discovering which binding reactions actually occur among thousands of candidate interactions between the interactor being characterized (either a protein, a domain or an antibody) and the molecular probes fixed to the glass slide. Most frequently, the majority of interactor-probe pairs do not interact and can be considered uninteresting with respect to the purpose of compiling a list of all targets recognized by a specific interactor molecule. Post-processing of the raw data provided by the scanner that measures the fluorescence intensity of the spots on the chip is required to discriminate interacting from non-interacting pairs (the post-processing procedure is described in greater detail in paragraph 4.1.1.). Analogously, Neural Network predictors can be used to score candidate target sequences according to their likelihood of being real biochemical ligands of the interactor (e.g. a particular domain) whose target recognition specificity is modeled by the predictor. High rank predictions may therefore be added to the list of targets as "predicted interactors", while low scoring predictions may be safely disregarded.

What has been just described is reflected in PepspotDB's data model in the classes derived from the abstract BinaryInteraction class, representing a generic binary (i.e. involving two and only two partners) interaction: ExperimentalInteraction and PredictedInteraction. The first class embodies those interactor pairs that have been experimentally observed to interact, whereas the latter holds the place of those interactor pairs that, according to

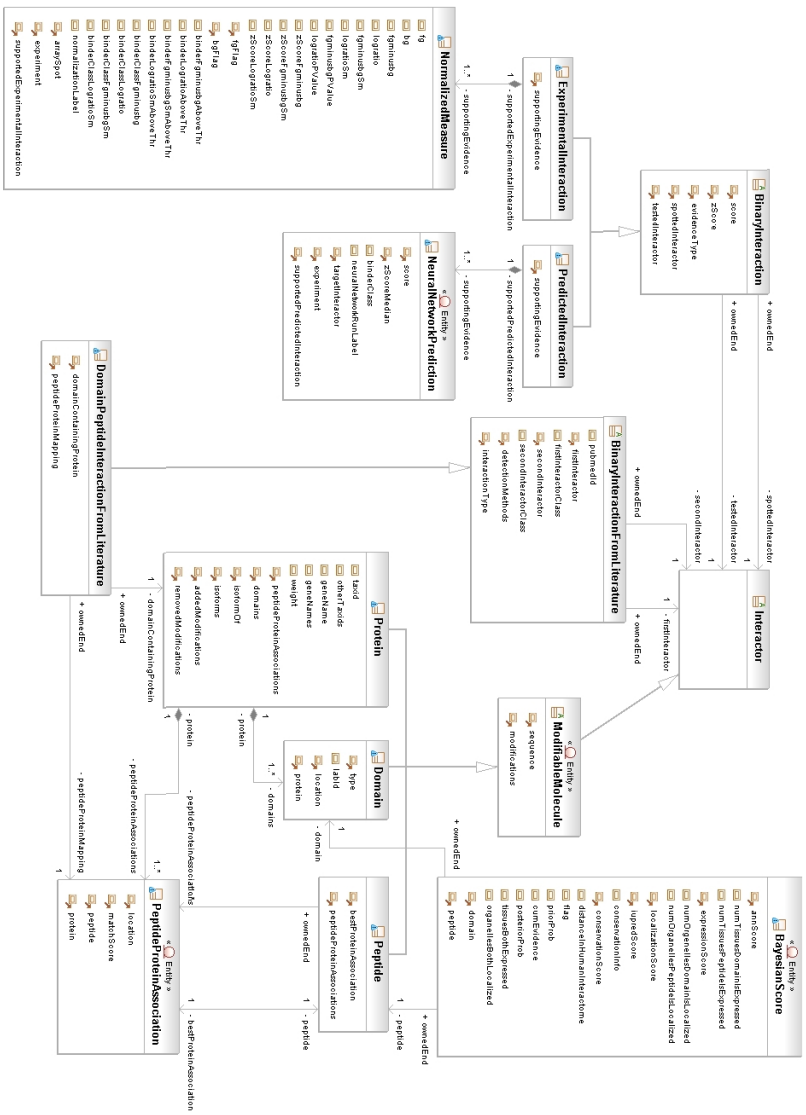


Figure 9: UML class diagram portraying the entities representing observed or predicted binary interactions, interaction data curated from the literature and Bayesian interaction confidence scores.

the judgment of a predictor trained on experimental data, are likely to interact actually. One may wonder why we deemed expedient to introduce new concepts in the model when the same information was already contained implicitly in the concepts modeling the results of both experiments and predictions (e.g. NormalizedMeasure and NeuralNetworkPrediction). The advantages of this approach are essentially two: 1) greater clarity and 2) better performance upon data retrieval. As for clarity, we regarded the mixture of "positive" (real interactions) and "negative" (non occurring) interactions in the rather generic concepts of "Measure" and "Prediction" as awkward and potentially confusing: thus, we singled out successfully detected interactions, creating a separate entity, BinaryInteraction, abstract common ancestor of further specialized concepts, ExperimentalInteraction and PredictedInteraction (see Fig. 9). A BinaryInteraction object is linked to the experimental or computational pieces of evidence that support it, as is apparent in the diagram. Secondly, speaking from a merely technical point of view, the distinction of "true", "positive" interactions from "negative" ones is even more called for: since candidate interactor pairs do not interact much more frequently than they do, if we stored all interaction data in the same table, the increasing amount of "negative" interaction data would certainly end up hindering the efficient interrogation and retrieval of truly interacting pairs, which are the only interesting ones for most biological inquiries.

Another separate kind of binary interaction is represented by the BinaryInteractionFromLiterature class. As stated by the name itself, a BinaryInteractionFromLiterature differs from a BinaryInteraction in that the information it contains was taken from already published articles and it is not necessarily supported by experimental observations or computational predictions stored in PepspotDB. The possibility to store in PepspotDB interaction data curated from the literature should not be intended as a viable alternative to more sophisticated solutions, such as MINT or IntAct, but rather as a minimal effort to address the needs of predictive algorithms,

which often require the assembly of high-confidence, specifically tailored golden standard sets of reliable interactions to serve as training data. This is exactly how this capability has been used so far: in the context of the SH2 human interactome mapping project, a golden standard set of experimentally observed interactions between human SH2 domains and target proteins, for which the particular phosphorylated tyrosine recognized by the domain, along with its flanking residues, was known, has been assembled, and domain-peptide interactions belonging to the golden standard set have been stored in the database as instances of DomainPeptideInteractionFromLiterature (a specialization of BinaryInteractionFromLiterature).

This "positive" golden standard set and a specular "negative" set were used to develop a Bayesian model, integrating different orthogonal information sources into a single score, rating the functional relevance of the candidate interaction. The integrated global scores produced by the Bayesian classifier have been also stored in the database (BayesianScore class, encapsulating all the information upon which the Bayesian inference is based), thus constituting a third type of interaction data (even though the term "interaction" only partially applies here: functional association would be more appropriate).

3.3.6. Controlled Vocabularies

When we confront the task of modeling any portion of reality, we often come across concepts whose properties cannot be adequately expressed with numeric attributes only: certain aspects of their nature denote quality, rather than quantity, and are therefore better captured by means of textual descriptions. True, many qualities may be categorized and in this way converted into an ordered series of integers, but, even though a computer

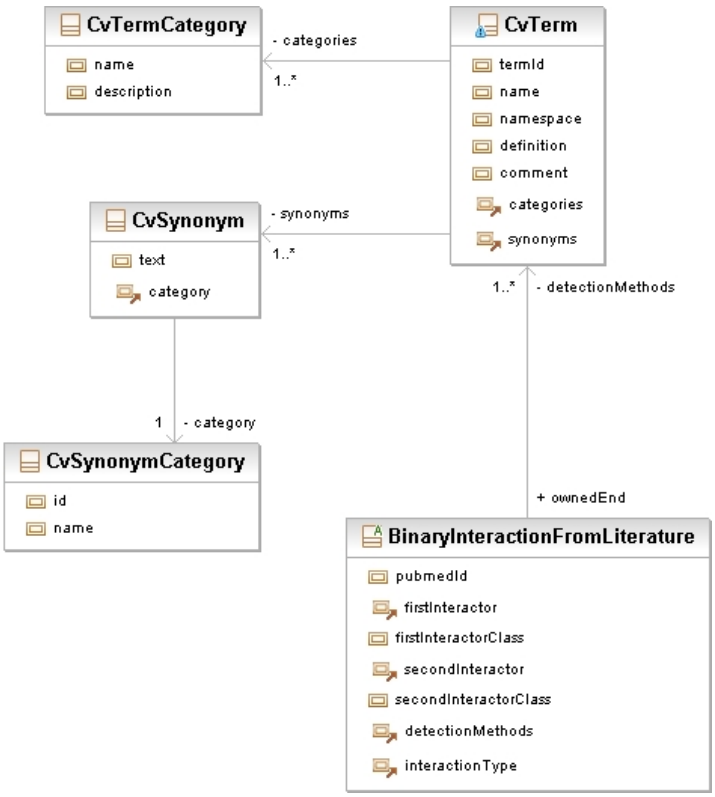


Figure 10: UML class diagram displaying the part of PepspotDB's data model implementing the concept of controlled vocabulary.

finds handling numbers more convenient and may well do so behind the curtains, in the end, categories should always be translated into meaningful phrases to facilitate human-machine interaction. Textual attributes present two inconveniences: 1) syntactic ambiguity; 2) semantic ambiguity. To put it

in simple words, without introducing extrinsic constraints to the form of expression and its content, we can never be completely sure that when different people use the same words, they mean the same things. This is where controlled vocabularies and ontologies come into the picture: a controlled vocabulary is in charge of establishing what words or phrases can licitly be assigned to a particular textual attribute; an ontology defines precisely and univocally the meaning of such words and phrases and the relationships between them. Controlled vocabulary and ontologies are a very effective means to enforce data consistency and to speed up searching, at the partial detriment of user creativity.

Compiling a controlled vocabulary is a difficult, time-consuming task, requiring great insight and expertise in the field. For small, pointed applications it may not be worthwhile to develop a brand new ad-hoc vocabulary. This is the reason why, in modeling textual attributes, in most cases, we were content with defining simple enumerations, without delving into the complex task of styling a full-fledged controlled vocabulary. Nevertheless, we included in PepspotDB's data model a few classes representing the concept of a controlled vocabulary: the class CvTerm and those directly associated with it (see Fig. 10). This capability allowed us to easily upload in PepspotDB the controlled vocabulary developed by the PSI-MI working group providing standardized annotations for molecular interaction data (the full vocabulary can be browsed online here: http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI). Molecular interactions extracted from the literature were then linked to the relevant interaction detection method annotations contained in the PSI-MI controlled vocabulary.

3.4. Web Interface

The object model that has been described in the previous paragraph was then translated into a relational structure: classes and relationships between classes were used as a blueprint to generate a classic relational database in tabular format where information could be stored permanently and manipulated on occurrence. Relational databases readily offer a powerful device for data management, interrogation and editing: virtually any operation on the data can be executed by issuing proper instructions, formulated in the SQL language, from a command-line terminal connected to the database server. Although interrogation by means of SQL queries may certainly sound appealing to computer programmers for its flexibility and power, the mere thought of writing correct statements in a programming language (even as simple as SQL) tends to send a shiver down the spines of wet-lab scientists, and rightly so. This motivated us to build on top of the relational database a user friendly web interface that would facilitate data access and retrieval to non computer experts. Albeit some desirable features and functionalities still remain to be implemented, the interface has reached sufficient maturity to be published on-line (as it will be soon). A description of the main sections the web site is divided in follows. Since only experiments aimed at discovering domain-peptide interactions have been loaded into the database so far, we will assume throughout the description to be dealing with experiments involving the probing of a domain with peptide arrays.

3.4.1. Homepage

Fig. 11 shows a screenshot of PepspotDB's web site. On the top right corner, below the title and top navigation bar, the login/logout form is

visible. Access to the web site is open to anyone using the "guest" account, which comes with reading privileges only. Authenticated users are granted different privileges in accordance with their assigned role(s): Reader, Writer, Curator or SuperUser.

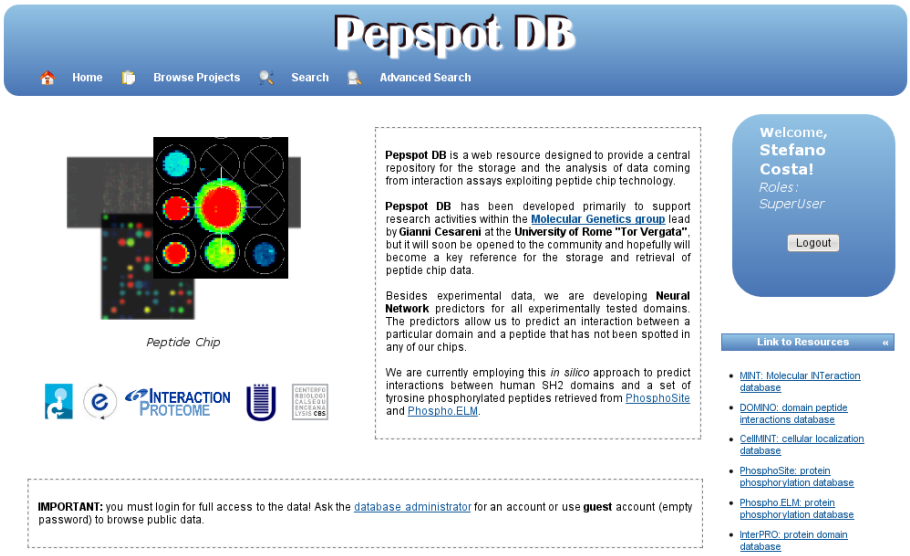


Figure 11: screenshot of PepspotDB's web site (http://mint.bio.uniroma2.it/PepspotDB).

3.4.2. Browse Projects

As was explained earlier, PepspotDB's short term goal was to support the research projects carried out in our lab, whose experimental part largely relied on interaction assays based on peptide array technology. Thus, an

entire section of the web site is dedicated to presenting these projects: the aim of each project is briefly described and its final results can be browsed through direct links. Fig. 12 shows a panel that summarizes the outcome of the SH2 Human Interactome Mapping project by showing a hierarchical clustering of the human SH2 domains according to sequence homology, where node color reflects the domain's target recognition specificity, as

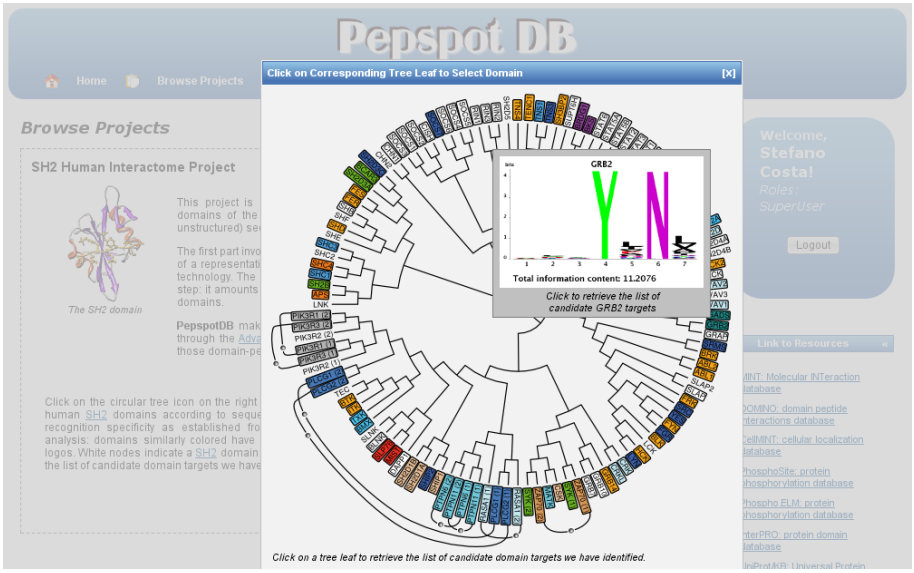


Figure 12: the "Browse Projects" section of PepspotDB's web site. The figure shows a panel that summarizes the outcome of the SH2 Human Interactome Mapping project by showing a hierarchical clustering of the human SH2 domains according to sequence homology, where node color reflects the domain's target recognition specificity. Domains similarly colored have similar consensus sequences; white nodes indicate a SH2 domain that we have not profiled yet. When the mouse cursor is dragged over a tree node, a sequence logo displaying the consensus sequence of the corresponding domain pops up (in the figure, GRB2 's logo is shown).

established from the results of our experimental and computational analysis. Domains similarly colored have similar consensus sequences, as is apparent from the logos popping up upon mouse hovering, whereas white nodes indicate a SH2 domain that we have not profiled yet. A click on a tree leaf forwards the user to the search page, where the candidate domain targets we have identified for that particular SH2 domain are listed.

3.4.3. Search

The Search page allows quick retrieval of domain-peptide interactions. After the protein containing the domain (let us call it protein A) and the protein containing the peptide (let us call it protein B) have been selected, interactions involving any of the domains belonging to protein A and/or (depending on user selection) any of the peptides belonging to protein B are found and are displayed on the page (see Fig. 13). It is also possible to specify a range within the sequence of the peptide containing protein, to narrow the result set to certain peptides only.

An interaction may be supported by experimental evidence, if it has been observed in at least one peptide chip experiment, by computational evidence, if it has received a sufficiently high score from a Neural Network predictor, or both. Which is the case can be easily grasped by looking at the search result: experimentally verified interactions are associated with an "experimental score", calculated as the logarithmic ratio between the foreground and background signals quantified by the scanner, whereas predicted interactions come with a "NeuralNetwork score", that is the output of the Neural Network predictor. Obviously, both scores are attached to interactions which are both experimentally and computationally supported.

On the bottom of the page there is a panel displaying several buttons controlling the set of operations that can be performed on the result set, such

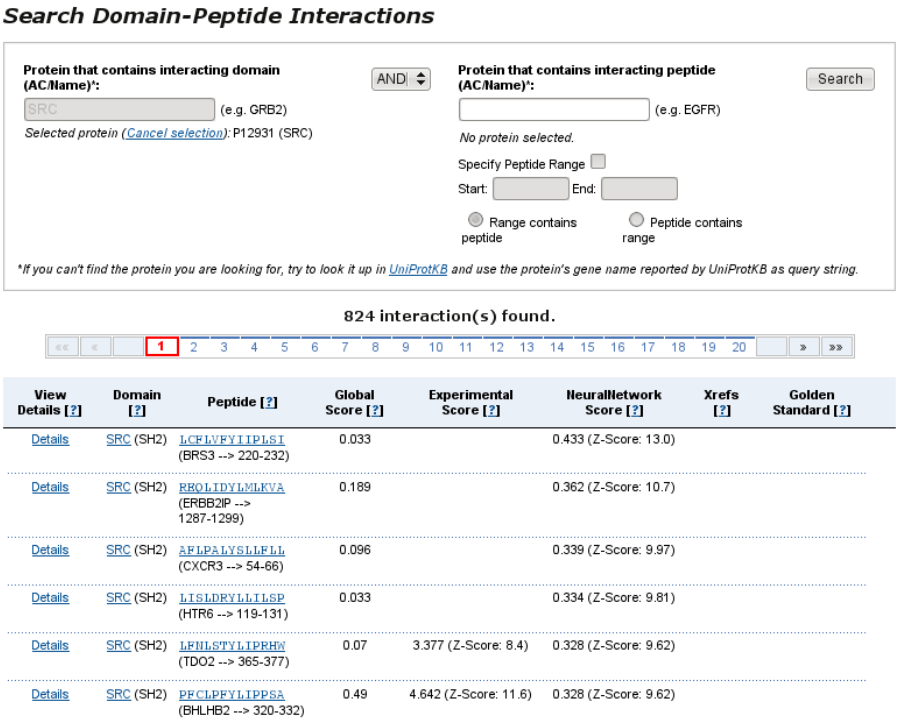


Figure 13: the "Search" section of PepspotDB's web site. The result of a query asking for interactions involving the SH2 domain of the SRC protein is shown. Each retrieved interaction is associated with up to three different scores: a "global score", calculated by the Bayesian classifier combining orthogonal sources of information, a "experimental score", calculated as the log-ratio of the foreground and background signals quantified by the scanner, a "Neural Network score", calculated by the Neural Network predictor. The global score is never missing, but either of the other two scores may be, in case the interaction has no support from experimental or computational evidence.

as filtering out records based on their content, sorting and exporting the query result in a textual format with comma-separated columns (Fig. 14).

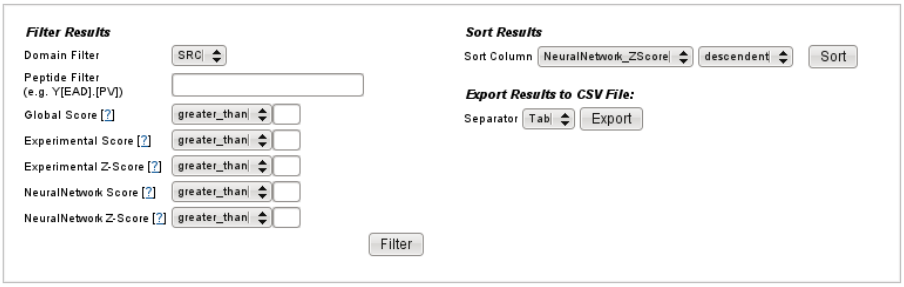


Figure 14: panel at the bottom of the "Search" page of PepspotDB's web site controlling the set of operations that can be performed on a query result set: filtering, sorting and exporting data to a text file (comma-separated columns).

Each retrieved interaction is also assigned a "global score", which is none other than the integrated score calculated by the Bayesian classifier combining together orthogonal sources of information. In addition to the score itself, it is also possible to inspect the pieces of contextual evidence that were combined in the Bayesian framework to obtain the final value. A simple click on "Details" opens up a new panel containing such information (Fig. 15).

Another valuable piece of information provided by the query result table regards the status of our previous knowledge of an interaction stored in PepspotDB: has an interaction between proteins A and B ever been observed by other groups in the world? The answer to this question is looked up in the main protein interaction database run by our group: MINT. Given a domain-peptide interaction between a domain of protein A and a peptide of protein B, if one or more protein interactions between A and B are found in MINT, regardless of whether the binding regions correspond or not, cross-references to the relevant MINT records appear on the correct line of the result table (Fig. 15). Furthermore, if PepspotDB contains one or more interactions, that

Figure 15: panel showing details about a domain-peptide interaction ("Search" page). The panel contains information on the experimental and computational evidence supporting the interaction and also on the contextual evidence that was combined by the Bayesian classifier to obtain a "global score" for the interaction. In the top right corner, links to relevant MINT records (mint leaf icon) and interactions curated from the literature (gold bar icon) are found.

have been manually curated into the database from the literature, for which both the proteins and the binding regions (e.g. domain and peptide) match, a

little gold bar icon shows up next to the cross-references. Clicking on the icon, a list of these "golden standard" interactions is produced, complete with links to the original papers they were taken from (Fig. 15).

3.4.4. Advanced Search

The search page that we have just described is very powerful for mining domain-peptide interactions, but PepspotDB allows the user to browse in great detail also proteins, domains, domain targets (e.g. peptides) and experiments. The "Advanced Search" page is the entry point to start digging

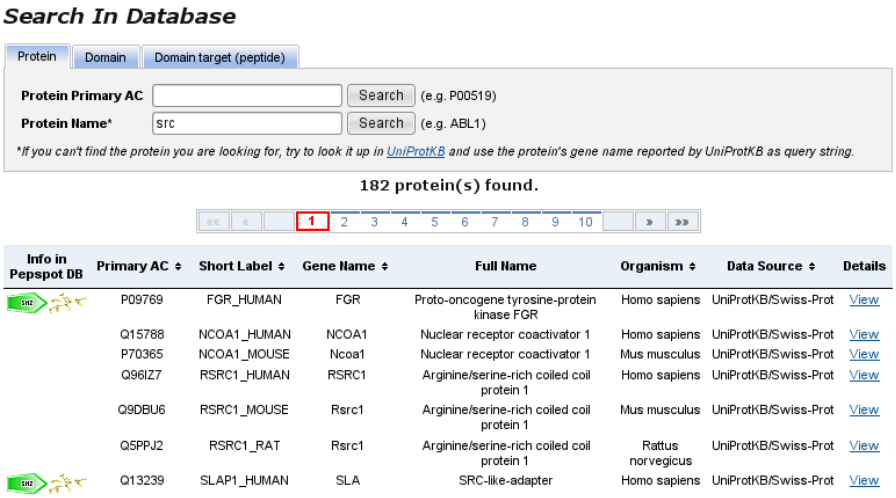


Figure 16: the "Advanced Search" section of PepspotDB's web site. The figure shows the result of a "search by protein name" query. The advanced search page allows the user to look for proteins, domains and domain targets (i.e. peptides) contained in the database. From there, one can proceed to exploring every detail about one's query hits either in the "Protein View", "Domain View" or "Peptide View" page.

into the available data (Fig. 16).

From there, we can move on to the "Protein View", the "Domain View"

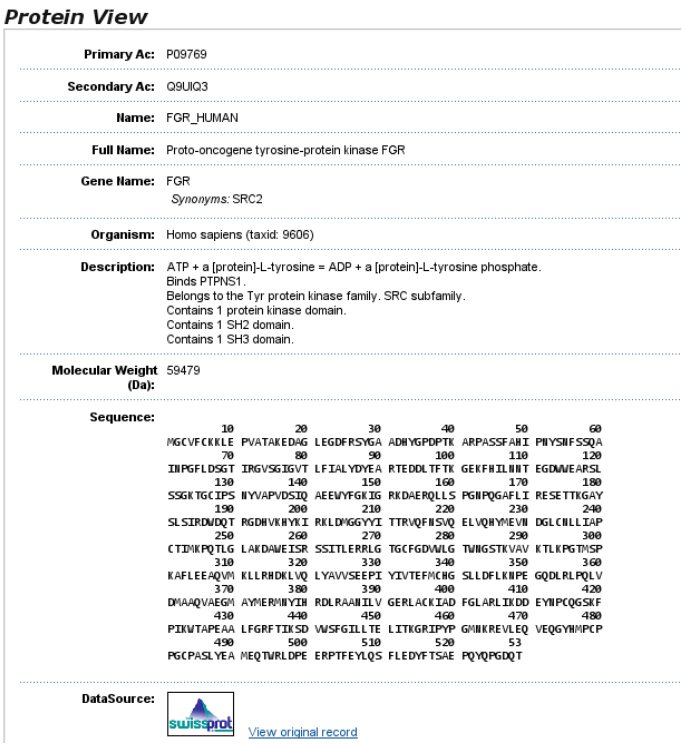
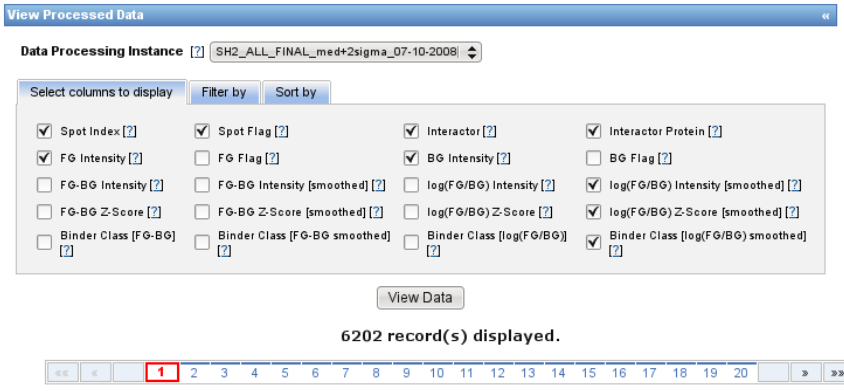


Figure 17: the "Protein View" page, providing a short description of the human kinase FGR (UniProt AC: P09769). The information has been imported from UniProtKB (http://www.uniprot.org; The UniProt Consortium 2007; The UniProt Consortium 2009).

or the "Peptide View", depending on the object of our quest. The "Protein View" provides us with a basic description of the selected protein, which is

essentially a short version of the UniProtKB description of the protein (Fig. 17). Moreover, we can find a list of post-translational modifications the protein may undergo carrying out its activity in the cell. The original source of the information is reported as well. At the bottom of the page, two panels display respectively a list of the peptides matching the sequence of the protein and a list of the domains the protein is composed of. It is important to note that in order to be listed here, a peptide or domain must have been



Spot Index	Spot Flag	Interactor	Interactor Protein	FG Intensity	BG Intensity	log(FG/BG) Intensity [smoothed]	log(FG/BG) Z-Score [smoothed]	Binder Class [log(FG/BG) smoothed]
2261	BAD_CYS	CTLWRAYANLHTA	BMX --> 143-155	44210.0	140.0	7.327	4.41	2
5185	GOOD	MSSLYYANALFS	HOXB7 --> 1-12	45656.0	145.0	7.318	4.4	2
1860	GOOD	RWLAVDYENVRPD	QAT --> 274-286	43768.0	146.0	7.258	4.37	2
6299	GOOD	DCAHFHYENVDFG	SNX26 --> 48-60	36668.0	141.0	7.03	4.22	2
4562	BAD_CYS	RVCRPCYRQLNRK	HGS --> 210-222	36395.0	143.0	6.996	4.2	2
1765	GOOD	DNDITPYLVSRFY	PRPF4B --> 843-855	35687.0	140.0	6.964	4.18	2
1849	BAD_CYS	WYRNCHYANLNGL	TNXB --> 4236-4248	35516.0	150.0	6.898	4.14	2

Figure 18: part of the "Experiment View" page, where the outcome of an experiment can be conveniently perused thanks to the possibility of selectively displaying, filtering and sorting the experiment's processed data.

tested in at least one experiment, or, in the case of a peptide, with at least one binding predictor.

The "Domain View" and the "Peptide View" have similar structure, with a general description at the top of the page and further details as we scroll down. The most relevant pieces of information the "Domain View" gives us are: 1) what experiments involving this domain are available; 2) what predictors have been trained for this domain. By clicking on the relevant links, we are taken either to the "Experiment View" or to the "Neural Network Predictor View", where the outcome of the experiment or of the predictor can be carefully scrutinized, manipulated through filtering and sorting, and finally exported (Fig. 18). There is also the possibility to draw a

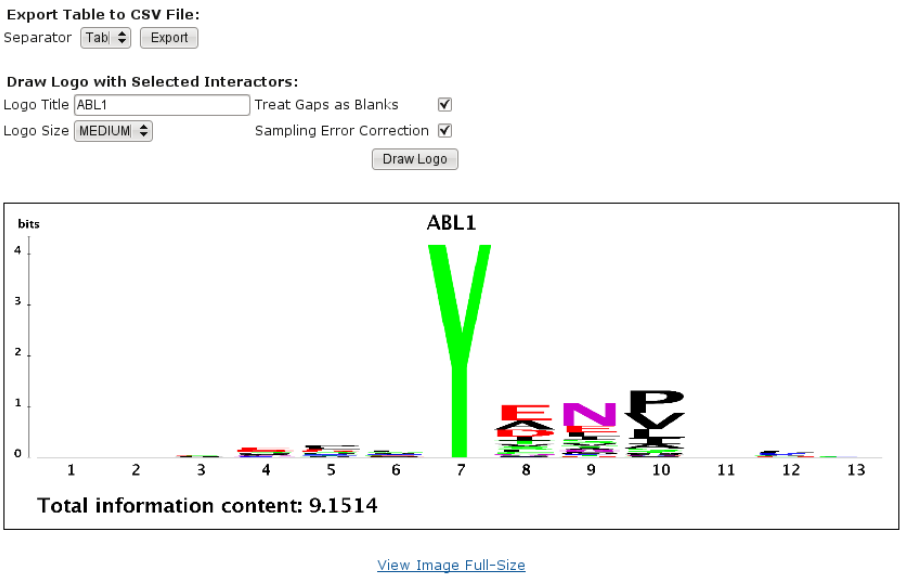


Figure 19: bottom part of the "Experiment View" page. The panel shown in the figure allows the user to export the selected data in a text file (comma-separated columns) and to draw a sequence logo using the sequences of the selected peptides.

sequence logo of the peptides currently selected and displayed in the table (Fig. 19). A logo is a graphical representation device that allows one to quickly grasp which positions, if any, in a multiple sequence alignment are enriched with particular amino acids, how much the information content of each position is and how much that of the whole alignment (Schneider et al. 1990).

The “Peptide View” collects four pieces of data about the selected peptide: 1) what protein sequences are matched by the peptide sequence (a range identifying the location of the match in the protein sequence is specified); 2) what modifications were effected on the peptide upon synthesis; 3) what experiments the peptide participated in and what was the outcome (observed to bind or not); 4) what predictors produced a score for the peptide and what these scores indicated (predicted to bind or not).

3.5. Software Used for PepspotDB's Development

The realization of PepspotDB required the development of multiple pieces of software employing different technologies and their concerted operation: 1) a relational database; 2) an object-oriented API implementing PepspotDB's data model and providing a low-level interface to populate the database tables, as well as to retrieve the data; 3) a web application providing a user-friendly, universally accessible, high-level interface to the data; 4) a collection of scripts to process experimental results and computational predictions; 5) a tool to draw sequence logos.

The first piece of software to have been developed is the objected-oriented API, which has been written in the Java language and has been built on top of the Java Enterprise platform version 5. The Enterprise Edition of the Java platform was chosen because the rich and advanced features of technologies like Enterprise Java Beans (EJB) 3.0, Java Persistence, Java

Transaction and Java Architecture for XML Binding (JAXB) 2.0 greatly facilitated the task of developing the server-side part of a distributed, transactional and data-oriented application. Thanks to the Java Persistence API, a Plain Old Java Object (POJO) model, implementing PepspotDB's data model, could be readily employed to generate a relational database schema. The underlying engine actually providing object-relational mapping (i.e. providing automatic conversion from Java objects to records in a relational database and viceversa) and querying services is Hibernate, probably the most powerful open source java persistence framework available.

The web application has been developed with JavaServer Faces (JSF) 1.2, which is also part of the Java Enterprise platform, a technology designed to simplify the building of user interfaces for JavaServer applications by providing a ready-to-use library of UI components with server-side event handling capabilities. To further reduce the complexity originating from the simultaneous employment of multiple advanced technologies, we exploited a powerful open source platform for building rich Internet applications in Java, called Seam. The Seam framework effectively glues together technologies such as Asynchronous JavaScript and XML (AJAX), JSF, Java Persistence and EJB 3.0: its unifying role is fundamental to simplify the development of web applications.

The tool for drawing sequence logos, dubbed rXLogo, has also been developed on the Java 5 platform; the Standard edition of the platform was used. The tool allows to draw a logo according to information content or relative entropy, to correct for sampling error, to calculate a frequency matrix from the alignment, to align the input sequences according to a user-defined regular expression and to produce multiple logos in a single run. Its source code has been integrated in PepspotDB's web application, though only a limited set of features are available in this online version.

The scripts used to process experimental results and computational

predictions have been developed with R, a language for statistical computing and graphics.

PepspotDB database runs on PostgreSQL 8.1, an open source Relational Database Management System (RDBMS), while PepspotDB web application runs on JBoss AS 4.2, a Java application server implementing the full J2EE 1.4 specification, plus some features of J2EE 5, such as EJB 3.0. Both server programs run on a dual-processor Intel Xeon 3.4 GHz machine, with 4GB RAM and two SATA 250GB Hard Drives configured in a RAID1 array.

4. Results

In this section we will describe the role played by PepspotDB and its companion tools in the SH2 Mediated Human Interactome Mapping Project, which, among the three large experimental projects currently being carried on in our lab, is the one that has reached the most advanced development stage. PepspotDB's contribution covers three main areas: 1) post-processing of raw experimental data; 2) selection of experimentally identified and predicted binders; 3) storage of raw data, processed data, experimental and predicted interactions and Bayesian integrated functional association scores.

4.1. SH2 Mediated Human Interactome Mapping Project

This project aims to the mapping of the SH2 mediated human interaction network. The project is now close to completion. The strategy we have employed comprises two main steps. The first part involves mostly wet lab work, since it consists of profiling the target recognition specificity of a representative collection of human SH2 domains (~70 out of ~110) by means of peptide array technology (Figure 20). The second part is essentially computational and relies on the data produced in the first step: it amounts to the development of neural network based predictors for each of the profiled SH2 domains. Of course, the experimental and computational aspects of this strategy are not independent, but rather tightly intertwined, and the outcome of the whole strategy depends on their mutual cooperation: an increase in the accuracy of the experimental dataset naturally triggers a performance boost in the predictors, and a good predictor may help to obtain an indication of the binding capabilities even of peptides for which, for various technical reasons, no clear answer could be gathered from the chip readout.

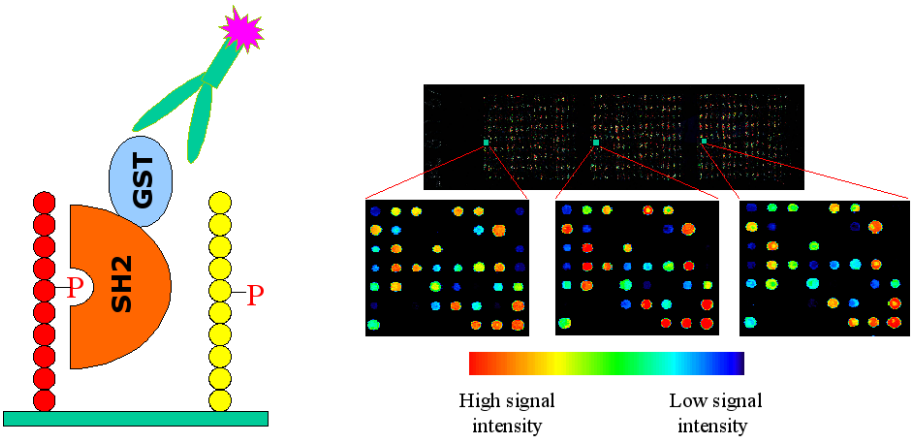


Figure 20: the figure shows the typical outcome of an experiment (in this case, the SH2 domain of the BRDG1 protein was profiled). On the left: the interaction between the SH2 domain, fused to GST, and a peptide spotted on the chip is revealed by an anti-pY antibody conjugated with a fluorophore. On the right: the same array of tyrosine-phosphorylated peptides is spotted in triplicate on a single microscope glass slide: this allows us to assess the reproducibility of the experiment.

The experiments were performed using the peptide array described in paragraph 3.1., containing approximately 6,000 distinct amino acid sequences spotted in triplicate. Thus, the final raw experimental dataset consists of 70 arrays (one for each successfully profiled domain) of ~18,000 measures each.

Our collection of 6,000 tyrosine phosphorylated peptides was a good representative of the entire human phosphotyrosine proteome when the project first started, but, as time went by, more and more phosphorylation sites were being discovered in the human proteome that were not represented in our array. This called for a complementary computational approach that

would enable us to estimate the likelihood of a sequence to bind to a domain, even though no direct biochemical measure of its affinity to the domain was available. Hence, we have developed a collection of Neural Network predictors that would learn from the experimental data a general model capturing the target recognition pattern of the domains and we have used them to: 1) correct potential experimental artifacts; 2) extend the collection of candidate binders by predicting on sequences that could not be tested experimentally. We have trained on the experimental dataset a total of 70 Neural Network predictors, one for each profiled domain. We employed Artificial Neural Networks of the standard three-layer feed forward type and encoded the amino acids as described by Nielsen et al. (2003). Only peptides with a length of 13 and with the phospho-tyrosine residue centrally placed were taken into account. To avoid over fitting the data set was homology reduced using CD-HIT (Li and Godzik 2006) with default values and 90% sequence identity threshold. These operations reduced the training data set for each SH2 domain from 6202 peptides to 3896. For each SH2 domain we normalized the log-ratio intensity values to range between 0 and 1, where higher numbers correlate with stronger binding affinity. The data set was divided into four subsets by random partitioning. We trained a Neural Network on two subsets, determined the optimal network architecture and training parameters on the third subset, and obtained an unbiased performance estimate from the fourth subset. This was repeated in a round-robin fashion to utilize all data for training, test, and validation. For each test set the number of hidden neurons in the ANN (0, 2, 4, 6, 10, 15, 20, and 30) were optimized according to the Pearson correlation coefficient. After training, each Neural Network predictor was used to obtain a binding prediction for each of the ~13,600 sequences resulting from the merging of the 6,000 sequences spotted on the chip with all the human 13-mers centered on a phosphorylated tyrosine residue, as reported by PhosphoSite (Hornbeck et al. 2004) and Phospho.ELM (Diella et al. 2004) databases. Neural

Network scores and signal intensity values show an average Pearson correlation coefficient of 0.4: this number is high enough to assure us that the predictors have indeed captured the general rules determining domain specificity, but at the same time it is low enough to indicate that no overfitting has occurred during the training phase.

4.1.1. Post-processing of Raw Experimental Data

To detect the peptide targets of each SH2 domain, we employed peptide array technology. The technique is conceptually simple: several thousand peptides are spotted on a microscope slide; then, the chip is incubated with a preparation of the domain fused to GST. Finally, the interaction between the domain and the peptide is revealed by an anti-GST antibody conjugated with a fluorophore: the chip is put in a scanner machine that excites the fluorescent epitopes with a laser and then quantifies the intensity of the signal emitted by each array spot. To identify the spots, the scanner software, prior to quantification, superimposes to the scanned image a grid where each spot occupies a circular area of fixed size and whose x-y coordinates have been previously defined. The average (median) intensity of the pixels laying within the area delimited by the circle gives the *foreground* signal of the spot; the average (median) intensity of the pixels falling outside of the circle, but still within a small distance from the circle's borderline, gives the *background* signal of the spot. The net signal intensity of a spot is computed as a function of the foreground and background signal, either by 1) subtracting background from foreground intensity (FG – BG); 2) taking the logarithm of the ratio of foreground to background intensity, which is equivalent to subtracting the intensity values in logarithmic rather than linear scale (log(FG/BG)). We considered both approaches and, in the end, opted for the latter, because it is more commonly employed in microarray analysis.

Raw data, namely the figures corresponding to the foreground and background signals of each spot in the array, may well suffer from experimental artefacts, like poorly phosphorylated peptides, bad grid positioning, smears, etc. The chip readout is also often affected by spatial biases and random noise. To cope with these sources of error, careful post-processing of the data is required.

The post-processing workflow involves several steps:

- *Removal of spots that do not react to anti-phosphotyrosine antibodies*
  In order to identify possibly flawed spots, one glass slide for each chip layout was probed against a collection of three different anti-pY antibodies, Tyr100, PY20 and 4G10. Firstly, the chips were tested separately with a preparation containing a single antibody, then, a mixture of all three antibodies was used. The spots which did not light up in either of the experiments (1,888 spots) were flagged 'BAD' and were not taken into account in further data processing steps.

- *Removal of spots containing overreacting Cys enriched peptides*
  An initial analysis of the peptides giving a strong signal in the GRB10, BTK, APS, SHD, TENC1, TXK, VAV1, SLP76 and TNS3 SH2 experiments showed that cysteine containing peptides had a propensity to bind non specifically to all domains. Although we have not clarified the reason for this aspecific binding we decided to exclude from the analysis 239 cysteine containing sequences.

- *Low signal correction*
  To prevent log-ratios from increasing indefinitely when background intensity is close to zero, we added a small fixed amount (delta) to all foreground and background intensity values. The value of delta is

defined on a per experiment basis and it is equal to the median background intensity of the experiment.

- *Assignment of 'high foreground' and 'low background' flags*
  In a few cases, due to a wrong positioning of the grid delimiting spot areas, the machine reads an unduly high background intensity value: in such situations, there is a risk of missing good binding candidates, namely spots with considerably high foreground intensity that would also have high FG-BG or log(FG/BG) values, if a mistake in background estimation had not occurred. To detect likely instances of this problem, we introduced two flags, 'high foreground' (fg_flag) and 'low background' (bg_flag): spots whose foreground intensity value is greater than two times the median foreground intensity of the experiment have their fg_flag set to 'GOOD'; conversely, spots whose background intensity value is greater than two times the median background intensity of the experiment have their bg_flag set to 'BAD'. Thus, problematic spots may be identified by looking for spots with fg_flag set to 'GOOD' and bg_flag set to 'BAD'.

- *Rescue of spots with unduly high background*
  Since it was impossible to discern automatically whether an unusually high background signal indicated a real faulty condition (e.g. dirty chip) or some recoverable artefact (e.g. error in grid positioning), we resorted to visual inspection to rescue some of the spots affected by the high-backgound problem and identified 34 instances that were clearly the result of a grid misplacement (see Figure 21). By setting the background signal of such spots to a more realistic value (the median background signal of the chip), we were able to detect a successful binding reaction in the subsequent analysis.

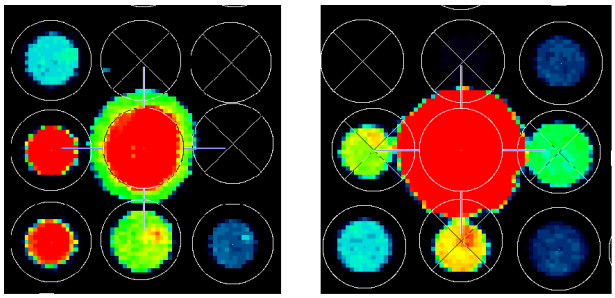


Figure 21: the figure shows, in the center of the two black rectangles, two spots whose background intensity is artificially high due to grid misplacement (the actual spot clearly extends beyond the circle's boundary). To correct the problem, the background intensity of both spots was set equal to the median background intensity of the experiment.

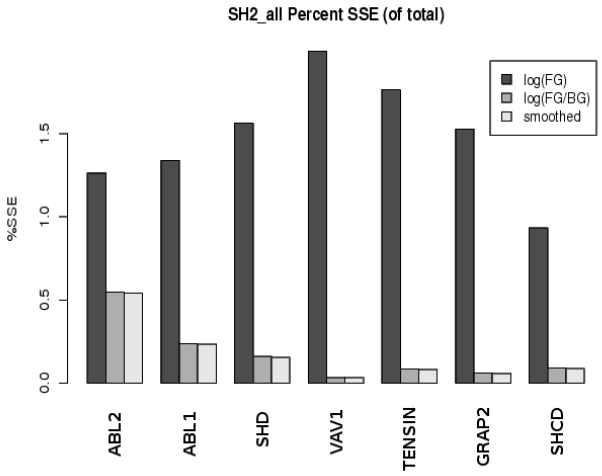


Figure 22: the figure shows the effect of smoothing on the Sum Squared Error (SSE), calculated from experiment replicates: the smoothing filter produces a slight SSE decrease, thus improving reproducibility.

- *Spatial smoothing*
  Microarray data is often affected by "spatial bias", i.e. certain areas of a chip look brighter or darker than others simply due to their position and not because of biologically relevant reasons. To cope with this problem, we applied a "smoothing" filter (Workman et al. 2002) that successfully removed some of the bias and mitigated random noise, as is shown by the increased reproducibility of the experiments (higher correlation between array replicates, see Figure 22).

- *Collapse replicates*
  Since experiments were performed in triplicate, the final foreground and background intensity values for each spot were computed by taking the median of the three replicated measures.

We used the average information content of sequence logos as benchmark to verify the improvement of the experimental dataset after post-processing (see Figure 23).

4.1.2. Selection and Storage of Experimentally Identified and Computationally Predicted Binders

Once post-processing had been completed, we could proceed to identifying the spots where a successful binding reaction had occurred. When the signal intensity of a spot exceeded a certain threshold, we considered that a successful binding reaction between the peptide spotted on that position in the chip and the tested domain had occurred. For each experiment, the threshold was set to the median signal intensity of the experiment plus two (or three) times the standard deviation from the median.

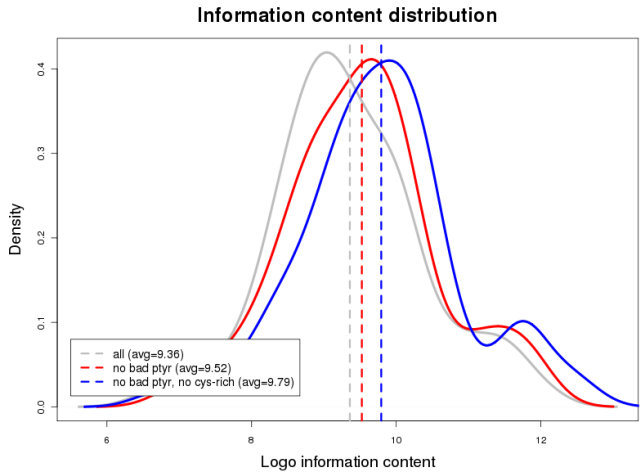


Figure 23: the figure shows the distribution of the total information content of the SH2 domains' logos obtained from the sequences of the peptides classified as "true binders", at three different stages of the post-processing workflow: prior to the removal from the dataset of any defective spot (grey line), after the removal of spots that did not react with anti-pY antibodies (blue line) and after the removal of spots that did not react with anti-pY antibodies or were enriched with cysteine residues (red line). The comparison supports the affirmation that successive steps of the post-processing workflow improve the quality of the experimental dataset, since the average logo information content increases accordingly.

Unfortunately, this operation is tainted with a certain degree of arbitrariness, since the choice of a threshold is always arbitrary, at least to a certain extent; yet, it is a reasonable solution that works well in most practical cases. Then, in order to favor spots with high foreground signal, we distinguished between three classes of binders: true binder, potential binder and non-binder, identified with 2, 1 and 0 respectively in the data files. True binders were defined as spots with signal intensity above the binding threshold and

having their fg_flag set to 'GOOD'. Spots with either signal intensity lower that the binding threshold or fg_flag set to 'BAD' were classified as potential binders. Finally, non-binders were those spots with both low signal and low foreground intensity (fg_flag = 'BAD').

A similar approach was employed to select "positive" binding predictions: sequences that had been assigned by the Neural Network predictor a score surpassing the median by more than two (or three) times times the standard deviation from the median were predicted to be true binders; if the score exceeded the median by a quantity between one and two times the standard deviation from the median, the sequence was predicted to be a likely binder; otherwise, the sequence was predicted to be a non-binder.

This simple strategy enabled us to obtain a reasonable average number of true binders per domain, summarized in Table 2.

Experimental		***Predicted***	
Threshold	Average number of binders	Threshold	Average number of binders
median+2*sigma	192	median+2*sigma	703
median+3*sigma	113	median+3*sigma	309

Table 2: the table displays the average number of experimentally identified and computationally predicted "true binders" for each SH2 domain, depending on the selected binding threshold (either median plus two times the standard deviation or median plus three times the standard deviation).

The consensus sequence preferentially recognized by each domain can be

visualized conveniently by drawing the logo of a domain's true binder sequences. Figure 24 compares some of the logos that we had experimentally identified with the motifs that were already known in the literature to encode the recognition pattern of the corresponding domains: notwithstanding minor discrepancies, most often the two representations agreed substantially.

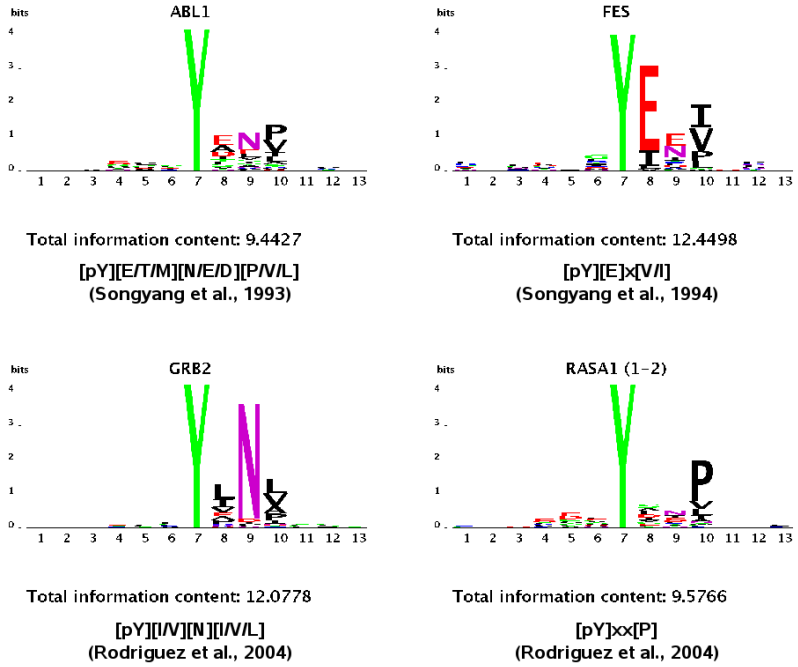


Figure 24: a comparison between four consensus sequences (displayed as sequence logos) that we have obtained from profiling experiments on the SH2 domains of ABL1, FES, GRB2, and RASA1 and the motifs that previous studies had used to encode the target recognition patterns of those domains. Notwithstanding minor discrepancies, the agreement between the two is quite good.

It is also interesting to compare the logos drawn from the experimentally determined true binders with those drawn from the computationally predicted true binders. In general, as Figure 25 shows, the letters in the logo encoding the predicted binders' consensus followed the same pattern as the logo encoding the experimentally determined binders' consensus, but were taller, seemingly indicating that the Neural Network predictor was successful in capturing the right consensus: sequences not strictly matching it were penalized and sequences strictly matching the consensus were rewarded (this

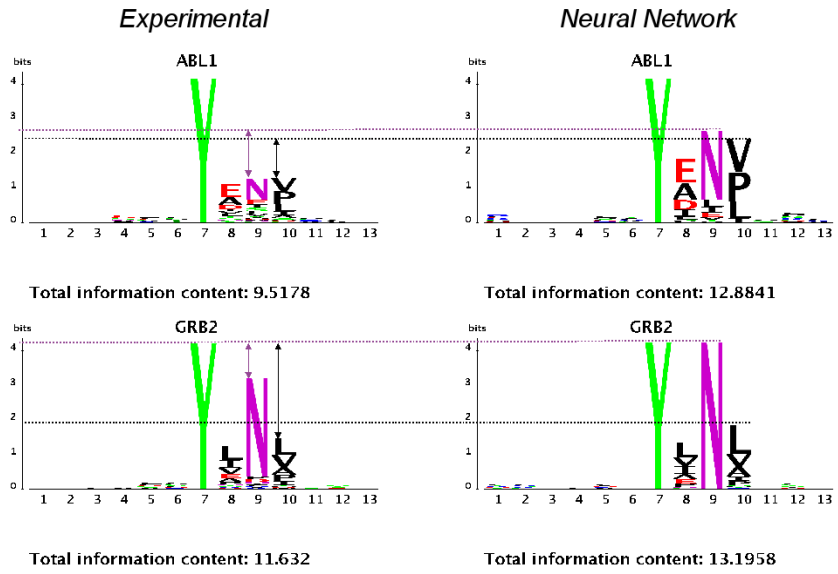


Figure 25: a comparison between the logos drawn from the experimentally determined true binders and those drawn from the computationally predicted true binders. The two series of logos are consistent, but the logos encoding the predicted binders' consensus generally have a greater information content.

would explain the increased average height of the letters). It should be emphasized here that, whereas Neural Networks can detect correlations between different sequence positions, logos are unsuited to visualizing them: the non-linear model learned by a Neural Network during the training phase is potentially more complicated and more informative than a logo can be.

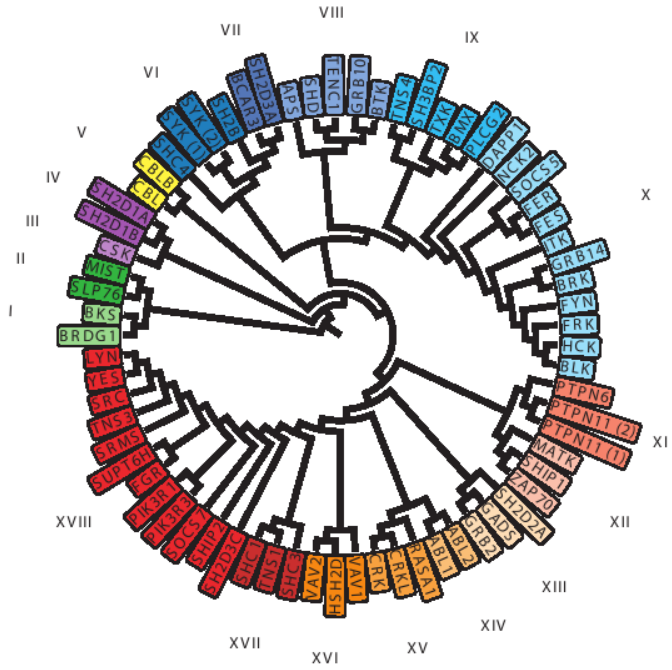


Figure 26: a hierarchical clustering of SH2 domains, where domains are clustered according to specificity. Nodes are colored according to the specificity class of the corresponding domain, indicated by the numbers surrounding the perimeter of the tree. We have defined eighteen distinct specificity classes.

The same software that was employed to draw the logos was also used to calculate the amino acid frequencies at each position from the set of 13-mers predicted by the Neural Network predictors to be true binders; these frequencies served to construct Position Specific Scoring Matrices (PSSMs), i.e. matrices where the generic element m_{ij} contains the frequency of amino acid j at position i. After having been converted into one-dimensional arrays, PSSMs were arranged so as to form the rows of a new matrix, a specificity matrix, where each row encoded the predicted target recognition pattern of a different SH2 domain. The specificity matrix was then used as input for a hierarchical clustering software that produced a tree-like structure where SH2 domains with similar specificity were grouped together under the same branch and SH2 domains with different specificity were separated into unrelated branches. By cutting the branches at a reasonable length, we defined a classification of SH2 domains into eighteen distinct specificity classes. Figure 26 shows a circular layout of the hierarchical clustering: nodes are colored according to the specificity class of the corresponding domain.

The logos and the clustering results clearly showed that SH2 domains have an intrinsic target recognition specificity. We then wondered to what extent similarities and differences in specificity could be correlated to similarities and differences in the SH2 domains' primary sequence. To investigate this point, we compared the hierarchical tree obtained using the specificity matrix as input with another tree where domains were clustered according to sequence homology (Figure 27). From the comparison, it is apparent that domains with related sequences, in not a few cases (e.g. CBL and CBLB, GRB2 and GADS, PTPN6 and PTPN11, BRDG1 and BKS and others), prefer target sequences matching similar patterns; however, some specificity groups exist (e.g. red, blue and light blue classes) whose members are found scattered over the entire tree, indicating that sequence homology is only partially predictive of specificity. Further studies, such as mutagenic

assays and bioinformatic analyses, are required to ascertain what residues are crucial to determine SH2 domain specificity.

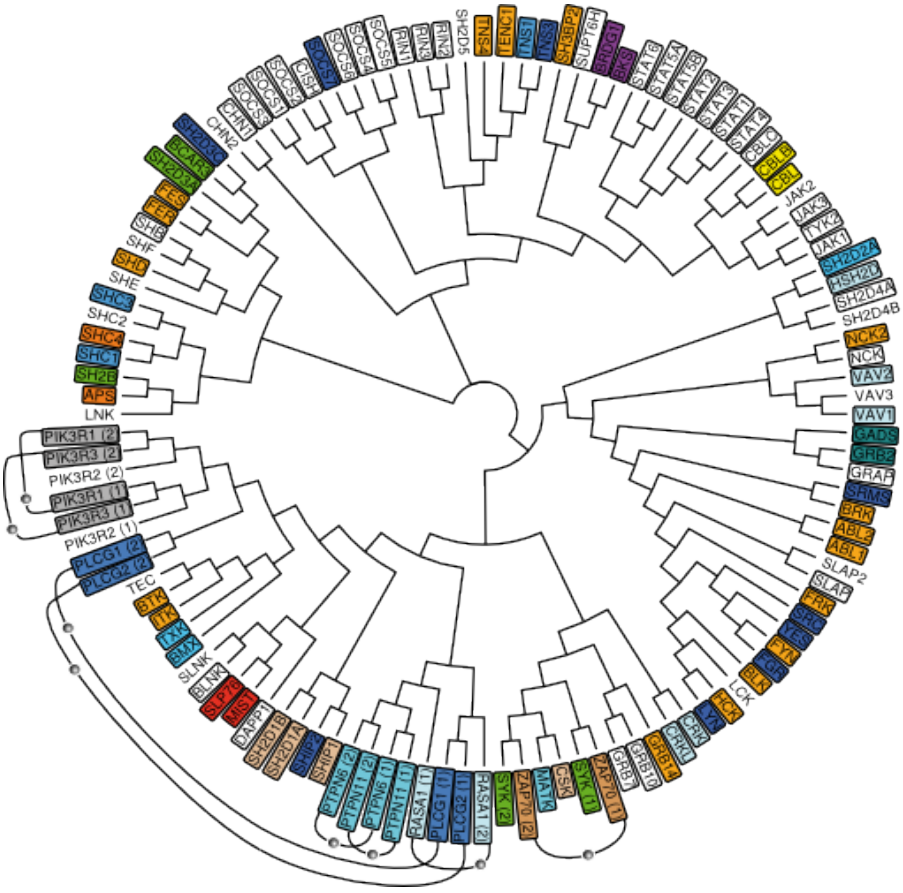


Figure 27: a sequence homology based hierarchical clustering of SH2 domains. Nodes are colored according to the specificity class of the corresponding domain; white nodes indicate SH2 domains that we have not profiled yet. The tree clearly shows that sequence homology is only partially predictive of specificity.

A total of 10,580 experimentally determined and 49,175 computationally predicted binary interactions involving 70 SH2 domains and 7,972 unique peptidic sequences were stored in PepspotDB. 4,207 interactions (~40% of experimentally determined and ~9% of computationally predicted interactions) are supported by both experimental and computational evidence. This can be considered a high-confidence set of SH2 domain-peptide interactions, as is indicated by its enrichment in domain-peptide interactions that were already know from the literature: 70 out of 276 (~25%, p-value < $1.11 \cdot 10^{-16}$).

4.2. Bayesian Scores Calculation and Storage

Since our experimental approach only provided us with *in vitro* evidence about whether a particular domain-peptide pair has the biochemical potential to interact physically or not, regardless of the physiological context where the interaction is to take place, we sought to complement our experimental observations with information that would help us to assess the functional relevance and the credibility of a candidate interaction, by taking into account currently established knowledge on several physiological conditions that may influence the supposed occurrence of the interaction. Given the diversity of the information that we wanted to combine, we carried through the integration task in the context of a Bayesian framework, where each feature is treated independently from the others and all features are seamlessly combined with the well-known Bayes' rule to calculate a global confidence score for the interaction (Figure 28).

The features that we have considered are:

- *Co-expression in tissues*: tissue-specific expression data were taken from Su et al. (2004); a co-expression score was obtained by counting the number of co-occurrences and dividing by the highest

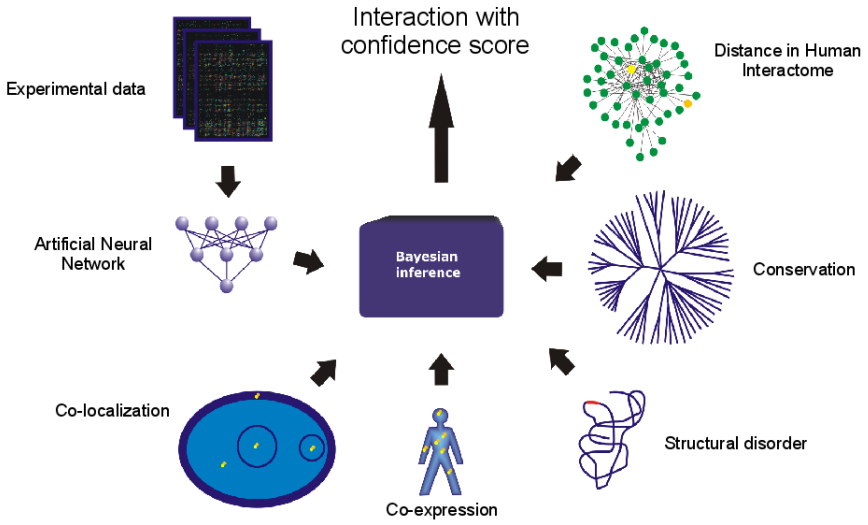


Figure 28: conceptual schema of our Bayesian framework, that seamlessly combines our Neural Network predictions with information about protein co-localization and co-expression, level of disorder in the protein containing the peptide, degree of conservation of the binding site in related species and distance between the supposedly interacting proteins in the human interactome, to obtain a single integrated interaction confidence score.

number of occurrences for either the SH2 domain containing protein or the peptide containing protein, thus obtaining a score between 0 an 1.

- *Co-localization in sub-cellular compartments*: sub-cellular localization data were extracted partly from CellMINT, a dedicated database currently being developed in our group (manuscript in preparation), and partly from GO annotations (Ashburner et al. 2000); a co-localization score was obtained by counting the number

of co-occurrences and dividing by the highest number of occurrences for either the SH2 domain containing protein or the peptide containing protein, thus obtaining a score between 0 an 1.

- *Structural disorder*: the level of structural disorder for the peptide containing protein was determined using IUPred by running the prediction method on the full sequences and then cutting out the relevant part (Dosztányi et al. 2005); a score between 0 and 1 was obtained by taking the average score of all the residues constituting the peptide.
- *Binding site conservation*: the degree of conservation of the binding site in related species was evaluated by inspecting it in multiple alignments of orthologs and paralogs from ENSEMBL (Hubbard et al. 2007). The relevant peptides were cut out of the related sequences and evaluated for binding by the neural networks. The score contribution for each orthologous sequence with the particular domain was calculated by multiplying the neural network score with the overall sequence distance from the original sequence obtained from a neighbor-joining tree. This procedure was followed to award binding site conservation in distant sequences more than that in close sequences. The scores obtained from all the orthologous sequences were summed to produce a single score for each binding site/SH2 domain combination: Cons. score = $\Sigma_i(dist_{sequence(i)} * ANN_{sequence(i)})$, where i runs through all orhologous sequences in the alignment for that particular peptide.
- *Distance in human interactome*: from a weighted human interactome (manuscript in preparation) network, we calculated the distance between the SH2 domain containing protein and the potential binding site containing protein using the Dijkstra algorithm. A higher score indicates higher distance in the interactome, which in turn reduces the chances that the SH2 domain containing protein and

the potential binding site come into contact.

- *Neural Network score*: Neural Network scores, transformed into Z-scores, were incorporated in the Bayesian framework as a feature on its own.

The Naïve Bayes rule is formulated as follows:

$$P(I|E) = P(I)*P(E1|I)*P(E2|I)...P(Ex|I)/\ P(E1)*P(E2)...P(Ex)$$

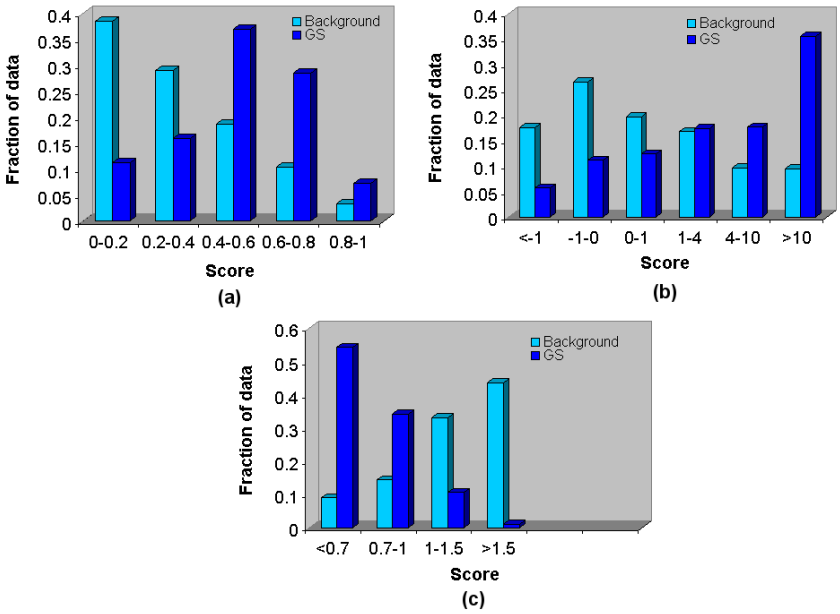


Figure 29: histograms showing the conditional probabilities of three features: (a) structural disorder; (b) binding site conservation; (c) distance in human interactome. Light blue bars indicate the probability of observing the evidence in the negative set (background set), dark blue bars indicate the probability of observing the evidence in the positive set (foreground or golden standard set).

This computes the probability of interaction given the evidence, P(I|E). The components of this calculation, P(Ex|I), are the probabilities of seeing each piece of evidence given interaction; P(Ex) is the probability of seeing this evidence in the full set of combinations of domain containing proteins and peptides. In practice, this latter probability is calculated by evaluating both the probability of the evidence given interaction and the probability of the evidence given non-interaction, i.e. P(Ex) = P(Ex|I) + P(Ex|NI).

The parameters for the model are determined from a set of 574 unique high-confidence SH2 domain-peptide interactions that were collected from the literature and curated manually (golden standard), deemed 'the foreground set', as well as the full range of possible combinations of SH2 domain containing proteins and peptides ('the background set'), assuming that most of these combinations are non-interacting *in vivo*.

The individual features used were transformed into probabilities by binning the feature score of both the foreground set and the background set. For any individual feature score, the probability of belonging to either set was then determined. Each feature was binned to maximize the difference between the foreground set and the background set while making sure each bin contained a reasonable number of samples. In the few cases this was not possible, a pseudo probability was used to avoid an excessive probability ratio between the foreground and the background probabilities. Figure 29 shows histograms representing the distribution of some of the features between the foreground and background set.

The efficiency of Bayesian scores in prioritizing golden standard interactions, as compared to experimental scores and Neural Network predictions, was evaluated by drawing ROC curves and calculating the Area Under the Curve (AUC). The results of this analysis for two different domains are displayed in Figure 30. In the case of PIK3R1, the Bayesian predictor clearly outperforms both the Neural Network predictor and the

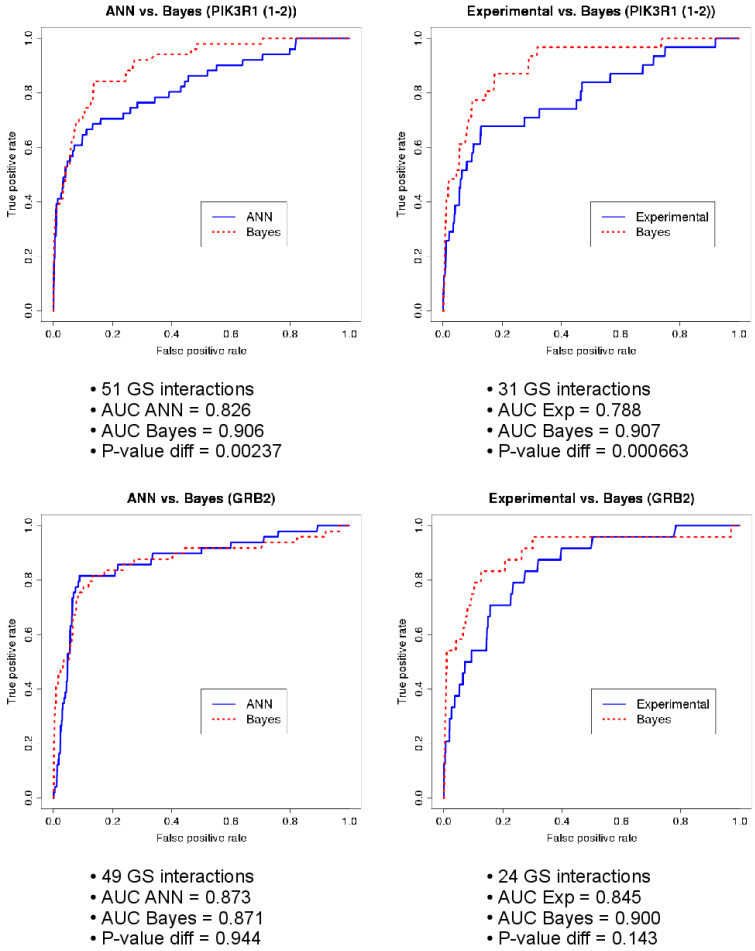


Figure 30: a comparison of the performance of the Bayesian classifier, the Neural Network and the processed experimental data, measured with respect to the golden standard set, for two different SH2 domains. In this type of graph, called Receiver Operator Curve (ROC), the true positive rate is plotted against the false positive rate. The larger the area under the curve, the better the performance of the classifier. An area greater than 0.9 indicates very good performance. In the case of PIK3R1, the Bayesian classifier clearly outperforms both the Neural Network and the experimental data, but in the case of GRB2 the curves cannot be distinguished. At the very least, the Bayesian classifier never performs worse than the others. However, the comparison is often vitiated by the drastically limited size of the golden standard set.

experimental data, whereas in the case of GRB2, the three ROC curves are not distinguishable. Although from these and other results it may be safely assumed that the Bayesian predictor never performs worse than the others, it is difficult to draw reliable conclusions: the number of golden standard SH2 domain-peptide interactions is limited and it is difficult to leave out a sufficient number of positive interactions to serve as an independent test set. This severely undermines the solidity of the analysis.

Bayesian scores were calculated for all possible SH2 domain-peptide pairs: a total of 955,010 scores were stored in PepspotDB, along with the information that was used to calculate the score.

5. Conclusions

5.1. What We Have Achieved

Nowadays it is generally acknowledged that small domains specifically binding to short linear peptides, whose amino acid sequences follow characteristic regular patterns, play a key role in a wide range of biological processes by mediating the assembly of stable and transient protein complexes, the latter being especially important for the correct operation of signal transduction pathways (Bhattacharyya et al. 2006; Pawson et al. 2001). Approximately three years ago, we set out to tackle the problem of mapping the target recognition specificity of several families of such domains, beginning with the human SH2 domain family, with a combined experimental and computational approach, exploiting both the recent advancements in peptide array technology and the predictive power of non-linear machine learning techniques. Furthermore, we have sought to complement our approach and assess the reliability of our inferences by integrating in a Bayesian model our results with all the available knowledge regarding the physiological conditions in which a candidate interaction ought to occur.

The considerably large scale of this effort immediately posed a big data management challenge: how could we effectively process the raw data, use the processed data to train Neural Network predictors and store all the results in a well-structured format, so as to facilitate further analyses and interrogations? No publicly available database designed to collect experiments based on peptide array technology existed at that time. The two databases of protein-protein interactions developed by our group, MINT and

DOMINO, could well have been used to store only the domain-peptide interactions that we would ultimately identify, but not to collect the entirety of the experimental and computational results, under pain of overcrowding the database with uninteresting (at least from the point of view of a user just looking for solidly established protein-protein interactions) data, with imaginable negative consequences on the database's performance.

Thus, we started the project which is the subject of this doctorate thesis, with the purpose of developing a new database-centered application, called PepspotDB, specifically designed as a repository for the storage and the analysis of interaction data coming from assays exploiting peptide array technology. A secondary objective was to lay the foundation for a centralized resource that could be made available to the community and become a reference site for the storage and retrieval of peptide chip data.

We started off by creating an object model, that is, a formal and unambiguous, albeit simplified, description of the portion of reality we desired to capture. The model has been implemented in the Java programming language and has been shaped into a library that henceforth has formed the core of the whole PepspotDB application. The Java object model has been used to create a traditional relational database, where all the raw data have been stored. On top of the core library, a rich web interface has been built, employing some of the most advanced technologies featured in the Java Enterprise platform, version 5. Additionally, a collection of tools for raw data processing has been developed separately. These tools have been devised to reduce background noise, to remove from the dataset any experimental artifacts (such as faulty spots or mistakes due to the scanner machine's software) and to automatically classify peptide probes in “true binders”, “likely binders” and “non-binders”. To display consensus sequences in a highly informative fashion, we have also developed a novel tool for drawing sequence logos, which has been integrated in the web application. Furthermore, we have trained Neural Network predictors (one

for each profiled domains) that have been used to predict binding for all tyrosine phosphorylated peptides which had been either spotted on our chips or stored in the PhoshoSite and Phospho.ELM databases. Finally, we have combined in a Bayesian framework our predictions with information about protein co-localization and co-expression, structural disorder, conservation of the binding site and protein-protein distance in the human interactome, to obtain a single integrated interaction confidence score.

The decision of developing a relational database *ex novo* granted us full freedom of action, allowing us to carefully tailor the database schema to the subject of our scientific inquiry, i.e. the storage and analysis of interaction data coming from assays employing peptide array technology. This custom design has been instrumental to achieve high performance even when dealing with huge amounts of data (PepspotDB currently counts more than 5 million records and this number is going to increase soon). At the same time, we have struggled to keep the database structure flexible enough to leave room for further expansion and improvement. In fact, PepspotDB can readily accommodate experiments based on any type of molecular array (peptide, protein, antibody, etc...): no modification to the schema is needed. Furthermore, thanks to PepspotDB's support for both deep and shallow integration strategies, the integration of external data sources with PepspotDB requires little additional effort.

PepspotDB comes with a rich web application providing a user friendly interface to the data: few clicks are necessary to execute sophisticated queries retrieving practically all the information available in the database. Query results can be edited directly on-line thanks to powerful filtering and sorting facilities, or can be downloaded in text format for further analysis. In the “Experiment View”, there is also the possibility to draw sequence logos on-the-fly from a group of selected peptides.

PepspotDB so far contains close to 80 experiments, from which we could successfully profile 70 human SH2 domains. Both the raw and processed

experimental data have been stored in the database. Besides the amino acid sequences spotted on the chips used for the experiments (~6,000), PepspotDB also contains all human peptides (13-mers) with a phosphorylated tyrosine residue in the central position, as reported by the PhosphoSite (Hornbeck et al. 2004) and Phospho.ELM (Diella et al. 2004) repositories of experimentally observed phosphorylation sites. For each of the profiled SH2 domains, a Neural Network predictor has been trained and applied to the full set of peptides collected in the database (~13,600 sequences); after processing, all the binding predictions were stored in the database as well. From the whole experimental and computational datasets, all binary domain-peptide interactions with strong support from either dataset (or both datasets) have been extracted: two sets of 10,580 experimentally determined and 49,175 computationally predicted binary interactions, with an overlap of 4,207 interactions, have thus been produced and stored in PepspotDB. Finally, Bayesian integrated confidence scores have been generated (one for each possible domain-peptide pair) and have been stored in the database. The total number of records in PepspotDB has by now surpassed the 5 million barrier and it is destined to grow, as soon as 20 experiments involving protein tyrosine phosphatases (PTPs) and 133 experiments involving domains binding proline-rich motifs enter the database.

In conclusion, we can say that we have succeeded in developing PepspotDB, a specialized, efficient and flexible new database for the storage and analysis of interaction assays based on peptide array technology. Its attached web interface facilitates the fruition of the data by the non-computer programmer and provides features that are useful to the biologist and the bioinformatician alike. The content of PepspotDB comprises the results of experimental observations as well as computational predictions, which complement and support each other. The inferred interactions are accompanied by contextual evidence that, integrated with the predictions in a

Bayesian confidence score, helps the user to assess the reliability and the biological relevance of an interaction.

We are confident that scientists struggling to untangle the web of protein interactions mediated by domains binding to short linear peptides will find the information and the tools provided by PepspotDB a useful resource to start formulating new testable biological hypotheses.

5.2. Future Developments

Happy as we are with the current development stage of PepspotDB, we are not willing to overlook the fact that there is still much room for improvement. Here are sketched a few guidelines for further developments, listed in order of priority:

- The tools for noise and artifacts removal should be rewritten in Java and integrated in the core library.
- The web application should be extended with a wholesale new section, “Data Upload”, through which database curators would be able to upload new data and edit already stored records.
- Automatic update procedures should be set up to bring records imported from external data sources in line with their current version in the original data source (especially urgent for UniProtKB).
- Access privileges should be more fully exploited, thus paving the way to a future opening of PepspotDB to external groups wishing to use it as a repository for their own experiments based on peptide array technology.
- rXLogo, the tool for drawing sequence logos, could be made available as a standalone application, or as an on-line tool accessible from the PepspotDB web site but not strictly dependent upon it.
- A dump of the entire SQL database should be available for

download; the Java code of both the core library and the web application should be publicly released.

5.3. Acknowledgements

Gianni Cesareni has coordinated the project.

Michele Tinti, Serena Paoluzi, Martina Carducci and Anita Palma have carried out the experiments with peptide arrays.

Emanuela Ferrari has thoroughly tested the web interface and has provided insightful advice for improving its usability.

Michele Tinti has manually curated the literature reporting SH2 domain-peptide interactions.

Lars Kiemer has developed the Bayesian integration framework.

Martin Lee Miller has developed the Neural Network software and has trained the predictors.

Chris Workman has devised the smoothing algorithm.

6. Appendix

6.1. Methods to Detect Protein-Protein Interactions

6.1.1. Two-Hybrid

The two-hybrid method is a genetic method that exploits the characteristic of some transcriptional factors of being composed of two domains, both essential for function, covalently bound: a N-terminal domain binds to specific DNA sequences and a C-terminal domain, negatively charged, necessary to activate transcription. None of the two domains expressed separately or co-expressed within the same cell can activate transcription. The principle of the two-hybrid method is based on the observation by Stanley Fields that the genetic fusion of the two domains composing the yeast transcription factor Gal4 to proteins interacting with each other results in a restored transcriptional activity, as revealed by the expression of a reporter gene. Apparently, the interaction between the proteins fused to the transcription factor's moieties is sufficient to bring the activation and binding domains so close to each other as to reconstitute a full functional transcription factor.

More generally, the two-hybrid method is based on the substitution of a covalent bond with a non-covalent bond mediated by the interaction of two distinct proteins. Besides Gal4, other transcription factors, e.g. LexA, or proteins of different function, like ubiquitine, β -galactosidase or β -lactamase have been used to design alternative two-hybrid methods.

The two-hybrid method can be applied in two different ways. One

possibility is to use a protein genetically fused to the transcription factor DNA binding domain as bait to fish out interactors from a clone mixture, obtained by fusing genetically the Gal4 activation domain to all the coding sequences of a specific proteome.

Another strategy, more apt to automation, involves the production of two separate collections of clones expressing hybrid proteins, fused respectively to the DNA binding domain and the activation domain. Then, all possible cross-combinations of the clones are assayed in a matrix format. This strategy allows to detect the interactions occurring between any pair of proteins in the proteome in a single experiment.

Advantages:

- Inexpensive
- High-throughput: DNA is manipulated, not proteins, so ORF libraries can be readily utilized
- Weak and transient interactions can be detected
- Independent of the abundance level of the endogenous proteins (proteins are overexpressed)

Drawbacks:

- Although the interaction occurs in a eukaryotic cell, two-hybrid cannot be considered an *in vivo* technique, because interactions are detected outside of their physiological context
- The interaction necessarily occurs in the nucleus, so it is difficult to study certain classes of proteins (e.g. membrane proteins)
- False positives may arise due to self-activation
- The detection of interactions dependent on some post-transcriptional modifications (e.g. phosphorylation) may be impaired due to the fusion of the protein to the transcription factor domain or due to the absence of the correct enzyme in the yeast nucleus.

6.1.2. Affinity Purification of Protein Complexes

This method has gained significant importance thanks to the development of advanced protein purification techniques through affinity chromatography and, most of all, of powerful instruments capable to measure with great precision even very limited amounts of protein. The approach starts by labeling a protein member of a complex with a tag that is recognized by a high affinity resin or a specific antibody. The components of the complex, co-purified with the tagged protein, are then identified by mass-spectrometry. This technology has been applied to the entire yeast proteome, allowing the retrieval of all the protein complexes in *S. cerevisiae*.

Advantages:

- Both binary and cooperative interactions can be retrieved
- Interactions are detected in their physiological context
- Only one member of the complex is expressed as a fusion protein, thus minimizing possible steric interferences due to the tag

Drawbacks:

- Weakly bound proteins may be washed off during the purification steps, thus giving rise to false negatives
- No clear indication about direct interactions among complex members
- The detection of interactions between proteins active in certain biological processes, such as transport or signaling, may be impaired due to technical difficulties arising in the purification of these particular classes of proteins
- Since the proteins forming the complex are expressed at endogenous concentration, some interactions may be missed simply because of the low abundance of the proteins involved in the interaction

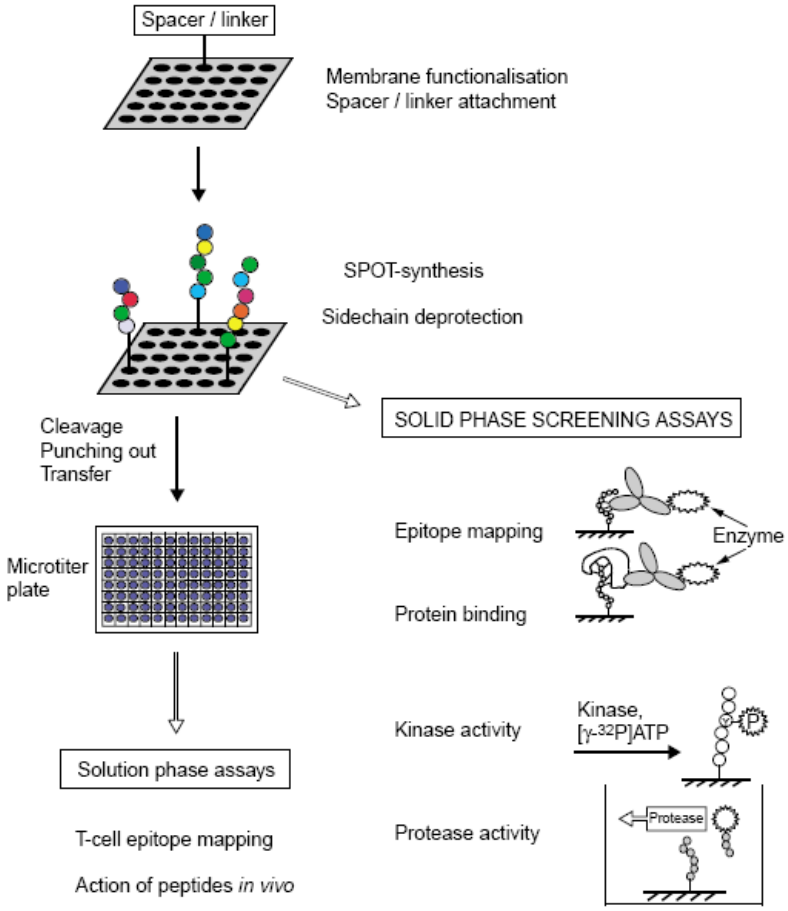


Figure 31: different applications of the SPOT-synthesis peptide arrays:

- epitope mapping: identification of peptides that promote immunological activity
- protein binding screening: detection of protein binding peptides
- enzyme-substrate screening: characterization of enzyme targets

(Reineke et al. 2001).

6.1.3. Peptide/Protein Array Technology

One of the most powerful *in vitro* techniques aiming at the systematic and, at least conceptually, simple discovery of all the interactors of a specific purified protein relies on peptide (or protein) array technology (Figure 31). The proteins of an entire proteome are expressed, purified and covalently bound to a solid support, like a cellulose membrane or a microscope glass slide, in an array layout.

Alternatively, short amino acid sequences can be directly synthesized on a solid support (usually cellulose) and arranged on a glass slide in a regular lattice. Then, the array is screened with a fluorescent probe (either another protein, a lipid, a peptide, or a domain): the fluorescence intensity of each spot, read by a laser scanner, reveals whether an interaction has occurred or not. This technique can be employed to study protein-protein, domain-domain, domain-peptide, domain-lipid or antibody-antigen interactions.

Advantages:

- Highly automatized
- The effect of post-translational modifications (e.g. phosphorylation) upon the interaction can be assessed and the substrates of different enzyme classes (e.g. kinases, phosphatases, metil-transferases, proteases) can be identified
- Experimental conditions can be carefully fine tuned: co-factors or inhibitors can be added, the intensity threshold above which an interaction is considered to have occurred can be properly established for each particular experiment
- Semi-quantitative technique
- Independent of the abundance level of the proteins in the cell, weak and unstable interactions can also be detected

Drawbacks:

- The interaction occurs outside of its physiological context
- High number of false positives
- Care is required for meaningful signal processing (background estimation, noise removal)

6.2. Public Databases Containing Protein-Protein Interactions

The deluge of protein interaction data following the widespread adoption of high-throughput interaction detection methods has prompted the rapid growth of molecular interaction databases, where interactions between proteins, nucleic acids and other organic molecules are stored, after careful curation by a team of experts, in a well-structured and well-organized fashion. Besides the data curation service, the groups responsible for developing and maintaining these databases often provide tools to help scientists with no knowledge about the database's internal structure to analyze and retrieve the data. Most databases are published on the Internet and are accessible through refined web interfaces enriched with complex features such as network visualization and analysis plugins.

Some of the most important publicly available repositories of protein interaction data: BIND (Biomolecular Interaction Network Database; Bader and Hogue 2000; Alfarano et al. 2005), DIP (Database for Interacting Proteins; Xenarios et al. 2000; Salwinski et al. 2004), BioGRID (General Repository for Interaction Datasets; Stark et al. 2006; Breitkreutz et al. 2008), HPRD (Human Protein Reference Database; Peri et al. 2003; Mishra et al. 2006), IntAct (Hermjakob et al. 2004a; Kerrien et al. 2007), MINT (Molecular INTeractions; Zanzoni et al. 2002; Chatr-aryamontri et al. 2007), and MPact (Güldener et al. 2006). The information gathered in these databases covers different kinds of molecular interactions (e.g. physical,

genetic or computationally inferred interactions), but the majority of the data is about physical interactions among proteins. The repositories differ between one another in coverage, data format, curation rules and degree of detail of the annotations. To cope with the challenge of capturing all the information available in the literature, five of the major interaction databases (BIND, DIP, IntAct, MINT and MPact) have come together in the IMEx (International Molecular Exchange) consortium, signing an agreement to share curation workload and exchange completed records on molecular interaction data (Orchard et al. 2007). The IMEx consortium represents a cooperative effort to provide the scientific community with a network of synchronized databases whereby all the available information on molecular interactions can be retrieved with little effort, similar to the successful global collaborations cataloging DNA sequences and solved protein structures. Data exchange will start soon. The consistency of all import-export operations involved in the data exchange process is guaranteed by the adoption of the PSI-MI (Proteomics Standard Initiative - Molecular Interaction) standard defined by the HUPO (HUman Proteome Organization) as common data model and exchange format (Hermjakob et al. 2004b).

7. Bibliography

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BFF, Hogue CWV (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33: D418--D424

Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7: 188–197

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM, Consortium I (2000) InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145--1150

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM,

Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-9

Bader GD, Hogue CW (2000) BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics 16: 465--477

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-42

Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H (2005) PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics 21: 827--828

Bhattacharyya RP, Reményi A, Yeh BJ, Lim WA (2006) Domains,motifs,and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem 75: 655–680

Blanchetot C, Chagnon M, Dubé N, Hallé M, Tremblay ML (2005) Substrate-trapping techniques in the identification of cellular PTP targets. Methods 35: 44--53

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294: 1351--1362

Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci 62: 435–445

Brannetti B, Zanzoni A, Montecchi-Palazzi L, Cesareni G, Helmer-Citterich M (2001) iSPOT: A Web Tool for the Analysis and Recognition of Protein Domain Specificity. Comp Funct Genomics 2: 314--318

Breitkreutz B-J, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res 36: D637--D640

Brooijmans N, Sharp KA, Kuntz ID (2002) Stability of macromolecular complexes. Proteins 48: 645–653

Castagnoli L, Costantini A, Dall'Armi C, Gonfloni S, Montecchi-Palazzi L, Panni S, Paoluzi S, Santonico E, Cesareni G (2004) Selectivity and promiscuity in the interaction network mediated by protein recognition modules. FEBS Lett 567: 74–79

Ceol A, Chatr-aryamontri A, Santonico E, Sacco R, Castagnoli L, Cesareni G (2007) DOMINO: a database of domain-peptide interactions. Nucleic Acids Res 35: D557--D560

Cestra G, Castagnoli L, Dente L, Minenkova O, Petrelli A, Migone N, Hoffmüller U, Schneider-Mergener J, Cesareni G (1999) The SH3 domains of endophilin and amphiphysin bind to the proline-rich region of synaptojanin 1 at distinct sites that display an unconventional binding specificity. J Biol Chem 274: 32001–32007

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572--D574

Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300: 1701-3

Consortium U (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res 35: D193--D197

Consortium U (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37: D169--D174

Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain-domain interactions from protein-protein interactions. Genome Res 12: 1540-8

Dente L, Vetriani C, Zucconi A, Pelicci G, Lanfrancone L, Pelicci PG, Cesareni G (1997) Modified phage peptide libraries as a tool to study specificity of phosphorylation and recognition of tyrosine containing peptides. J Mol Biol 269: 694–703

Diella F, Cameron S, Gemünd C, Linding R, Via A, Kuster B, Sicheritz-Pontén T, Blom N, Gibson TJ (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 5: 79

Diella F, Gould CM, Chica C, Via A, Gibson TJ (2008) Phospho.ELM: a database of phosphorylation sites--update 2008. Nucleic Acids Res 36: D240--D244

Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433--3434

Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21: 410--412

Frank R (2002) The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports: principles and applications. J Immunol Methods 267: 13–26

Freund C, Kühne R, Park S, Thiemke K, Reinherz EL, Wagner G (2003) Structural investigations of a GYF domain covalently linked to a proline-rich peptide. J Biomol NMR 27: 143–149

Gong W, Zhou D, Ren Y, Wang Y, Zuo Z, Shen Y, Xiao F, Zhu Q, Hong A, Zhou X, Gao X, Li T (2008) PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. Nucleic Acids Res 36: D679--D683

Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, Stümpflen V (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436--D441

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SGN, Sander C, Bork P, Zhu W, Pandey

A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol 22: 177--183

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32: D452--D455

Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 4: 1551--1561

Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. Nucleic Acids Res 35: D610--D617

Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J Mol Biol 362: 861–875

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35: D561--D565

Lacount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE (2005) A protein interaction network of the malaria parasite Plasmodium falciparum. Nature 438: 103-7

Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol 2: E14

Lee H, Deng M, Sun F, Chen T (2006) An integrated approach to the prediction of domain-domain interactions. BMC Bioinformatics 7: 269

Li L, Wu C, Huang H, Zhang K, Gan J, Li SSC (2008) Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. Nucleic Acids Res 36: 3263--3273

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658--1659

Liu J, Rost B (2004) CHOP proteins into structural domain-like fragments. Proteins 55: 678–688

Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HGM, Nagini M, Kumar GSS, Jose R, Deepthi P, Mohan SS, Gandhi TKB, Harsha HC, Deshpande KS, Sarker M, Prasad TSK, Pandey A (2006) Human protein reference database--2006 update. Nucleic Acids Res 34: D411--D414

Mongioví AM, Romano PR, Panni S, Mendoza M, Wong WT, Musacchio A, Cesareni G, Fiore PPD (1999) A novel peptide-SH3 interaction. EMBO J 18: 5300--5309

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJA, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. Nucleic Acids Res 35: D224–D228

Ng S-K, Zhang Z, Tan S-H (2003) Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19: 923-9

Ng S-K, Zhang Z, Tan S-H, Lin K (2003) InterDom: a database of

putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 31: 251--254

Nye TMW, Berzuini C, Gilks WR, Babu MM, Teichmann SA (2005) Statistical analysis of domains in interacting protein pairs. Bioinformatics 21: 993-1001

Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31: 3635--3641

Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Nerothin J, Hermjakob H (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. Proteomics 7 Suppl 1: 28--34

Pagel P, Oesterheld M, Stümpflen V, Frishman D (2006) The DIMA web resource--exploring the protein domain network. Bioinformatics 22: 997--998

Paoluzi S, Castagnoli L, Lauro I, Salcini AE, Coda L, Confalonieri SFaS, Pelicci PG, Fiore PPD, Cesareni G (1998) Recognition specificity of individual EH domains of mammals and yeast. EMBO J 17: 6541–6550

Pawson T (2007) Dynamic control of signaling by modular adaptor proteins. Curr Opin Cell Biol

Pawson T, Gish GD, Nash P (2001) SH2 domains,interaction modules and cellular wiring. Trends Cell Biol 11: 504–511

Pawson T, Raina M, Nash P (2002) Interaction domains: from simple binding events to complex cellular behavior. FEBS Lett 513: 2–10

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363--2371

Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein domain interactions. Nucleic Acids Res 36: D656--D661

Rain JC, Selig L, Reuse HD, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schächter V, Chemama Y, Labigne A, Legrain P (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409: 211-5

Reineke U, Volkmer-Engert R, Schneider-Mergener J (2001) Applications of peptide arrays prepared by the SPOT-technology. Curr Opin Biotechnol 12: 59--64

Rickles RJ, Botfield MC, Weng Z, Taylor JA, Green OM, Brugge JS, Zoller MJ (1994) Identification of Src,Fyn,Lyn,PI3K and Abl SH3 domain

ligands using phage display libraries. EMBO J 13: 5598–5604

Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. Genome Biol 6: R89

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449--D451

Santonico E, Castagnoli L, Cesareni G (2005) Methods to reveal domain networks. Drug Discov Today 10: 1111-7

Schmidt EE, Davies CJ (2007) The origins of polypeptide domains. Bioessays 29: 262–270

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097--6100

Scott JK, Smith GP (1990) Searching for peptide ligands with an epitope library. Science 249: 386–390

Shoemaker BA, Panchenko AR, Bryant SH (2006) Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein Sci 15: 352–361

Sparks AB, Quilliam LA, Thorn JM, Der CJ, Kay BK (1994) Identification and characterization of Src SH3 ligands from phage-displayed random peptide libraries. J Biol Chem 269: 23853–23856

Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers

of protein-protein interaction. J Mol Biol 311: 681–692

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535—D539

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062—6067

Sudol M (1998) From Src Homology domains to other signaling modules: proposal of the'protein recognition code'. Oncogene 17: 1469–1474

Vaccaro P, Brannetti B, Montecchi-Palazzi L, Philipp S, Citterich MH, Cesareni G, Dente L (2001) Distinct binding specificity of the multiple PDZ domains of INADL,a human protein with homology to INAD from Drosophila melanogaster. J Biol Chem 276: 42122–42130

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild H-H, Nielsen C, Brunak S, Knudsen S (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol 3: research0048

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28: 289--291

Yaffe MB, Cantley LC (2000) Mapping specificity determinants for protein-protein association using protein fusions and random peptide

libraries. Methods Enzymol 328: 157–170

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