doi:10.1006/jmbi.2001.4920 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **311**, 681–692

JMB

Correlated Sequence-signatures as Markers of Protein-Protein Interaction

Einat Sprinzak and Hanah Margalit*

Department of Molecular Genetics and Biotechnology The Hebrew University - Hadassah Medical School POB 12272 Jerusalem 91120, Israel

**Corresponding author*

As protein-protein interaction is intrinsic to most cellular processes, the ability to predict which proteins in the cell interact can aid significantly in identifying the function of newly discovered proteins, and in understanding the molecular networks they participate in. Here we demonstrate that characteristic pairs of sequence-signatures can be learned from a database of experimentally determined interacting proteins, where one protein contains the one sequence-signature and its interacting partner contains the other sequence-signature. The sequence-signatures that recur in concert in various pairs of interacting proteins are termed correlated sequence-signatures, and it is proposed that they can be used for predicting putative pairs of interacting partners in the cell. We demonstrate the potential of this approach on a comprehensive database of experimentally determined pairs of interacting proteins in the yeast *Saccharomyces cerevisiae*. The proteins in this database have been characterized by their sequence-signatures, as defined by the InterPro classification. A statistical analysis performed on all possible combinations of sequence-signature pairs has identified those pairs that are over-represented in the database of yeast interacting proteins. It is demonstrated how the use of the correlated sequence-signatures as identifiers of interacting proteins can reduce significantly the search space, and enable directed experimental interaction screens.

Keywords: protein-protein interaction; functional genomics; proteomics; bioinformatics; sequence-signature

Introduction

Protein-protein interaction is intrinsic to most cellular processes in the cell, whether in the form of multi-subunit proteins or as transient complexes of proteins. Therefore, unraveling which proteins in a cell work together is essential to the understanding of its functional networks. Until very recently the identification of interacting proteins was done exclusively in the laboratory, using a variety of experimental methods. In most cases, a particular molecular system was addressed and the study focused on specific proteins. Currently, the completion of many genome projects has enabled the development and application of new experimental approaches that attempt to discover the relationships between genes at a genomic scale.[1–6] One method that is widely used for large-scale protein-protein interaction analysis is the two-hybrid system.[7,8] Recently, large-scale two-hybrid system analyses were performed on several complete genomes,[7] including the yeast *Saccharomyces cerevisiae*,[9,10] vaccinia virus,[11] hepatitis C virus,[12] *Helicobacter pylori*,[13] and proteins involved in vulval development in *Caenorhabditis elegans*.[14] These experiments provide rich data resources of interacting proteins.

Recent computational approaches take advantage of the availability of full genomic information in an attempt to discover relationships between genes. These approaches do not rely directly on sequence similarity between genes (or their products), but rather use different types of information to reveal the functional relationships between molecules.[15,16] For example, a systematic comparison of the gene order among different genomes has revealed that in most cases the gene order is not conserved except for gene pairs, whose products usually involve interacting proteins.[17] Other studies have shown that pairs of proteins in a given genome that were fused into a single protein in another genome are likely to be related.[18,19] In another study it was demonstrated that proteins

E-mail address of the corresponding author:
hanah@md2.huji.ac.il

that function together have evolved in a correlated fashion in different genomes.[20] The many known relationships that were identified in these studies supported the conjecture that new relationships can be inferred from such analyses. Such approaches have led to the prediction of hundreds of pairs of related proteins in the *Escherichia coli* genome, and thousands in the genome of *S. cerevisiae*.[18,21] In general, it could be shown that integration of the predictions obtained by the different computational approaches together with experimental data, such as that provided by RNA expression patterns, can improve the functional assignments, as demonstrated for the yeast *S. cerevisiae* genome.[21]

The above approaches use mostly sequence considerations to point to a possible link between proteins, but do not use the available experimental data of interacting proteins. A major challenge is to learn from such data what typifies the sequences of interacting protein pairs, and to use this knowledge in prediction. This is the focus of the current study. We demonstrate that characteristic pairs of sequence-signatures can be learned from a database of experimentally determined interacting proteins, where one protein contains the one sequence-signature and its interacting partner contains the other sequence-signature. The sequence-signatures that appear together in interacting protein pairs more often than expected at random are termed correlated sequence-signatures. In the current study the terms significant sequence-signature pairs, over-represented sequence-signature pairs and correlated sequence-signatures are used interchangeably.

We apply such an analysis to a database of experimentally identified interacting protein pairs in *S. cerevisiae*. For typifying the proteins in this database by their signatures we use classifications supplied by the InterPro database. This is an integrated database that assigns to protein sequences typical signatures in the form of regular expressions, profiles, fingerprints and hidden Markov models, compiled from various databases.[22] Although the data of interacting proteins are relatively sparse, distinct over-represented sequence-signature pairs could be pointed out. Their power as predictors was demonstrated by a cross-validation test, supporting the promise of this approach. We propose that the correlated sequence-signatures can be used as markers for predicting new, as yet unknown, relationships between pairs of proteins, where each protein is characterized by one of the signatures. Although not every protein with the one sequence-signature is expected to interact with every protein with the other sequence-signature, these identifiers can be used to direct the experimental interaction screens, reducing significantly the search in the space of protein pairs.

Results

The identification of correlated sequence-signatures in interacting proteins involves several steps, as depicted in Figure 1. (i) Generation of a non-redundant database of experimentally determined interacting protein pairs to be used as a learning set. (ii) Characterization of the protein sequences by their sequence-signatures, based on the InterPro classification. (iii) Derivation of a contingency table of sequence-signatures (signatures by signatures). (iv) Identification of over-represented sequence-signature pairs by comparing their observed frequencies to those expected at random. Here we describe the potential of this approach by applying it to a data set of yeast interacting proteins.

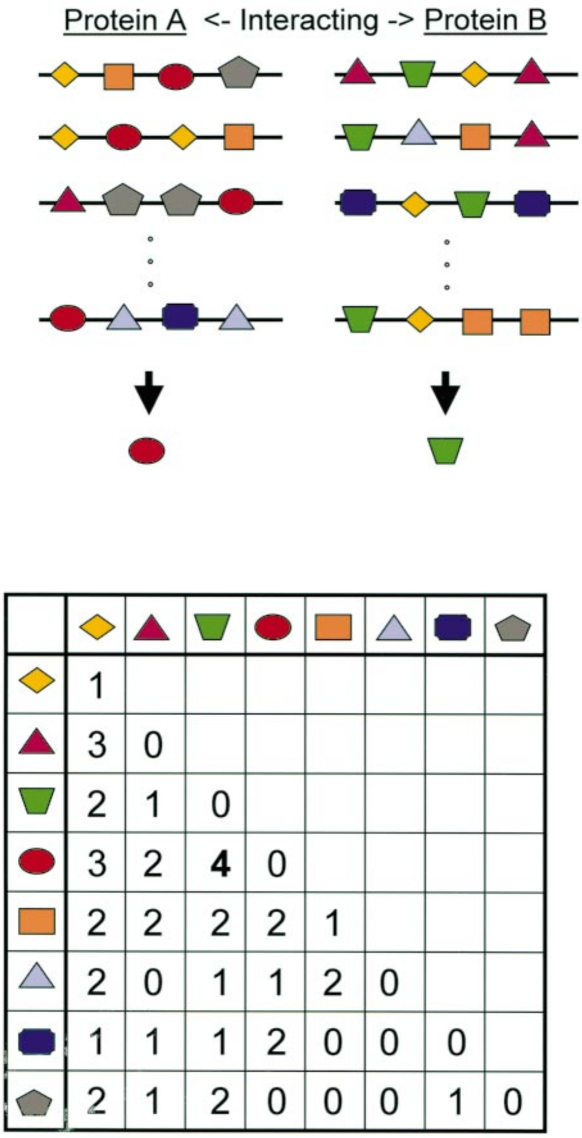
Construction of the contingency table of sequence-signatures

A non-redundant database of all experimentally defined pairs of interacting proteins in the yeast *S. cerevisiae* was constructed as described in Methods. This database included 2908 protein pairs. Next, the proteins were classified by InterPro signatures[22] (see Methods). Since annotation by InterPro signatures is not available for all proteins, this reduced the database of protein pairs to 1274 pairs, where both pair-mates had InterPro signatures. Based on this database, a contingency table of sequence-signatures was derived. We counted every combination of two signatures that appeared in the pairs of interacting proteins, where one sequence-signature appeared in one protein and the other sequence-signature appeared in its interacting partner. A signature that appeared several times in a protein sequence was recorded only once, in order to prevent a bias in sequence-signature pairs composed of motifs that tend to recur along protein sequences (Figure 1).

In total, the proteins in the database were characterized by 434 sequence-signatures. These were used for the generation of a 434 × 434 contingency table. The sum of all entries in the table is 2286, as several proteins contributed more than one signature to the table; e.g. if one protein in a pair contained signature a and the other pair member contained signatures b and c, this protein pair added to the count of both entries a,b and a,c. As expected for such a large table and such a small data set, most of the table entries remained empty: 1433 entries contained counts above zero.

Identification of over-represented sequence-signature pairs

To find the pairs of sequence-signatures that were more frequent than expected at random, we calculated the ratio between the observed and expected frequencies for each entry in the table and expressed it in log-odds values. The log-odds value was computed as $\log_2(P_{ij}/P_iP_j)$, where P_{ij} is the observed frequency of a sequence-signature



	◆	▲	▼	●	■	△	⬣	⬟
◆	1							
▲	3	0						
▼	2	1	0					
●	3	2	**4**	0				
■	2	2	2	2	1			
△	2	0	1	1	2	0		
⬣	1	1	1	2	0	0	0	
⬟	2	1	2	0	0	0	1	0

Figure 1. A schematic representation of the analysis for detecting correlated sequence-signatures in interacting proteins. In the upper panel, each row contains the sequences of a pair of proteins (A,B) whose interaction was determined experimentally. Each sequence is characterized by its signatures, denoted here schematically by colored shapes. In the lower panel, a contingency table of the signature combinations is described, where each entry (*i,j*) in the table shows the number of protein pairs that contain signatures *i* and *j* in concert, where one protein contains signature *i* and its pair mate contains signature *j*. For example, the sequence-signature pair represented by an orange rectangle and a pink triangle appears in two pairs of interacting proteins. The most abundant pair of sequence-signatures is that of a red ellipse and a green trapezium that appears in four different pairs of interacting proteins. In the next step of the analysis the likelihood of the identified sequence-signature pairs is evaluated.

pair (*i,j*) and P_i and P_j are the frequencies of sequence-signatures *i* and *j* in the data, respectively. The computed log-odds values ranged between −1.95 and 12.16. When the observed frequency was zero, the log-odds value was assigned to −2.5 (smaller than the minimum value obtained, which was −1.95). The information content of the table (the average mutual information per entry) was calculated by:

$$H = \sum_i \sum_j P_{ij} \log_2(P_{ij}/P_i P_j)$$

Despite the sparseness of the data, the information content was 2.48 bits. Thus, the information about a characteristic sequence-signature in one protein reduces the uncertainty about its potential interacting partners in 2.48 bits of information, on average.

The full list of signature pairs, their counts, and log-odds values can be found at http://bioinfo.md.huji.ac.il/marg/Int-signature. Favorable sequence-signature pairs can be identified by their positive log-odds values: 1141 pairs of sequence-signatures had log-odds values of at least 2 (i.e. the observed frequency was at least four times greater than that expected at random). There were many entries that got high log-odds values, although their actual counts were very low. This was due to the low frequency of many of the signatures in the database, implying that the expectation of two sequence-signatures to appear in concert was even lower. Therefore, even a count of 1 for a pair of sequence-signatures may be meaningful. However, until more cases are collected, the biological significance of these sequence-signature pairs remains questionable. The more interesting cases involve pairs of sequence-signatures with both relatively high counts and high log-odds values, and as more data of interacting proteins are accumulating these are expected to increase. In the following analysis we concentrated on sequence-signature pairs with log-odds values greater than or equal to 2 and counts greater than or equal to 5. There were 40 sequence-signature pairs above these thresholds (Table 1).

Over-represented sequence-signature pairs

In the following sections we describe in detail several examples of sequence-signature pairs from Table 1, and discuss various implications that can be deduced from the data presented in the Table. We use the sequence-signature names and accession numbers (IPRxxxxxx) as they are noted in the InterPro database. Signatures that were defined by the InterPro database as related were clustered and re-named (*CLUS00001*, *CLUS00002*, etc., see Methods and Table 2). In most cases this grouped signatures that describe a family of proteins and its sub-families.

Figure 2 summarizes the results for the yeast interacting proteins that contain the pair of signatures of the *Myb* domain (IPR001005) and the *Bromodomain* (IPR001487). As illustrated in the Figure, among the proteins that exhibit the *Myb* domain, this is the only signature that is shared by all three

Table 1. List of sequences-signature pairs with counts ⩾ 5 and log-odds value ⩾ 2

				Yeast proteins with signature	
Log-odds value	Counts	Sequence-signature I[a]	Sequence-signature II[a]	I[b]	II[b]
2.52	21	IPR000553 Cyclin	*CLUS00003*[c]	16	234
5.11	18	IPR001163 Sm protein	IPR001163 Sm protein	16	16
2.08	12	IPR000961 Protein kinase C-terminal domain	*CLUS00003*[c]	10	234
7.43	10	IPR001830 Trehalose-6-phosphate synthase	IPR001830 Trehalose-6-phosphate synthase	5	5
6.39	10	IPR000038 Cell division GTP binding protein	IPR000038 Cell division GTP binding protein	7	7
2.21	10	IPR000651 Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	*CLUS00006*[c]	6	75
2.18	10	IPR001895 Guanine-nucleotide dissociation stimulators CDC25 family	*CLUS00006*[c]	7	75
2.09	10	*CLUS00007*[c]	*CLUS00006*[c]	23	75
2.9	9	IPR000504 RNA-binding region RNP-1 (RNA recognition motif)	IPR000504 RNA-binding region RNP-1 (RNA recognition motif)	56	56
2.79	9	IPR001452 SH3 domain	IPR001452 SH3 domain	25	25
6.77	8	IPR002161 Uncharacterized protein Family UPF0030	IPR001852 Uncharacterized protein Family UPF0019	3	3
6.57	8	IPR002161 Uncharacterized protein Family UPF0030	IPR003009 Proteins binding FMN and related compounds	3	11
5.66	8	IPR000608 Ubiquitin-conjugating enzymes	IPR000608 Ubiquitin-conjugating enzymes	16	16
3.59	8	IPR001410[d] DEAD/DEAH box helicase	IPR001650[d] Helicase C-terminal domain	60	68
6.56	7	IPR000862 Replication factor C conserved domain	IPR000862 Replication factor C conserved domain	9	9
5.87	7	IPR001388 *Synaptobrevin*	*CLUS00008*[c]	5	23
5.18	7	*CLUS00008*[c]	*CLUS00008*[c]	23	23
4.31	7	IPR000629[d] ATP-dependent helicase, DEAD-box	IPR001410[d] DEAD/DEAH box helicase	22	60
4.10	7	IPR001440 TPR repeat	IPR001440 TPR repeat	36	36
3.79	7	IPR000629[d] ATP-dependent helicase, DEAD-box	IPR001650[d] Helicase C-terminal domain	22	68
2.62	7	IPR001680 G-protein beta WD-40 repeats	IPR001680 G-protein beta WD-40 repeats	92	92
7.35	6	*CLUS00025*[c]	IPR000842 Phosphoribosyl pyrophosphate synthetase	15	15
6.84	6	IPR002863 MutS family, N-terminal putative DNA binding domain	IPR002099 DNA mismatch repair proteins mutL/hexB/PMS1	5	4
6.74	6	IPR000432 DNA mismatch repair protein MutS family	IPR002099 DNA mismatch repair proteins mutL/hexB/PMS1	6	4
6.45	6	*CLUS00010*[c]	*CLUS00010*[c]	19	19
5.77	6	IPR001852 Uncharacterized protein Family UPF0019	IPR003009 Proteins binding FMN and related compounds	3	11
5.57	6	IPR001939 AAA-protein	IPR001939 AAA-protein	29	29
2.36	6	IPR001895 Guanine-nucleotide dissociation stimulators CDC25 family	IPR001452 SH3 domain	7	25
2.07	6	IPR001849 PH domain	IPR001452 SH3 domain	27	25
7.14	5	*CLUS00018*[c]	*CLUS00018*[c]	63	63
6.19	5	IPR000938 *CAP-Gly* domain	*CLUS00002*[c]	4	11
5.48	5	IPR001752 Kinesin motor domain	IPR001752 Kinesin motor domain	6	6
4.44	5	IPR002004 Poly-adenylate binding protein, unique domain	IPR000504 RNA-binding region RNP-1 (RNA recognition motif)	1	56
4.05	5	IPR001005 *Myb* DNA binding domain	IPR001487 *Bromodomain*	19	10
4.03	5	IPR001440 TPR repeat	IPR001404 Heat shock hsp90 proteins	36	2
3.12	5	IPR000690 RNA-binding protein C2H2 Zn-finger domain	IPR000504 RNA-binding region RNP-1 (RNA recognition motif)	7	56
2.61	5	*CLUS00009*[c]	IPR000934 Serine/threonine specific protein phosphatase	17	22
2.55	5	IPR001660 SAM domain	*CLUS00006*[c]	3	75
2.13	5	IPR000651 Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	IPR001452 SH3 domain	6	25
2.01	5	*CLUS00007*[c]	IPR001452 SH3 domain	23	25

The Table is sorted by the number of signature pairs (count). The full list of all yeast sequence-signature pairs can be found at: http://bioinfo.md.huji.ac.il/marg/Int-signature

[a] The signature accession number and its description according to the InterPro database.

[b] Number of proteins in the yeast genome that are predicted to contain the sequence signature.

[c] The description of the sequence-signatures included in a cluster is given in Table 2.

[d] Sequence signatures IPR000629, IPR001410 and IPR001650 are very similar, but still are defined separately by the InterPro classification.

Table 2. Signature clusters

CLUS No.	InterPro acc.[a]	Signature name[b]
CLUS00002	IPR003008	Tubulin/FtsZ family
	IPR000217	Tubulin family
	IPR002452	Alpha tubulin
	IPR002453	Beta tubulin
CLUS00003	IPR000719	Eukaryotic protein kinase
	IPR002290	Serine/threonine protein kinases active-site
	IPR001245	Tyrosine kinase catalytic domain
CLUS00006	IPR002380	Transforming protein P21 RAS
	IPR001806	Ras family
	IPR001609	Myosin head (motor domain)
	IPR000795	GTP-binding elongation factor
	IPR000251	ADP-ribosylation factors family
	IPR002046	SAR1 GTP-binding protein family
	IPR002041	GTP-binding nuclear protein Ran family
CLUS00007	IPR000963	LTE1/rasGRF-associated domain
	IPR001926	Beta family of pyridoxalphosphate dependent enzymes
	IPR001216	Cysteine synthase/cystathionine beta-synthase P-phosphate attachment site
	IPR000634	Serine/threonine dehydratase pyridoxal-phosphate attachment site
CLUS00008	IPR000017	Syntaxin/epimorphin family
	IPR000727	t-SNARE coiled-coil domain
	IPR000928	SNAP-25 family
CLUS00009	IPR002048	EF-hand family
	IPR001125	Recoverin
CLUS00010	IPR002423	TCP-1/cpn60 chaperonin family
	IPR002194	Chaperonins TCP-1
	IPR001844	Chaperonin cpn60 (60kDa subunit)
CLUS00018	IPR001066	Sugar transporter
	IPR000803	Facilitated glucose transporter family
CLUS00025	IPR000836	Phosphoribosyl transferase
	IPR002375	Purine/pyrimidine phosphoribosyl transferase

The list contains only signature clusters that appear in Table 1. The full list of 31 yeast signature clusters can be found at: http://bioinfo.md.huji.ac.il/marg/Int-signature.

The scheme for clustering of signature is described in Methods.

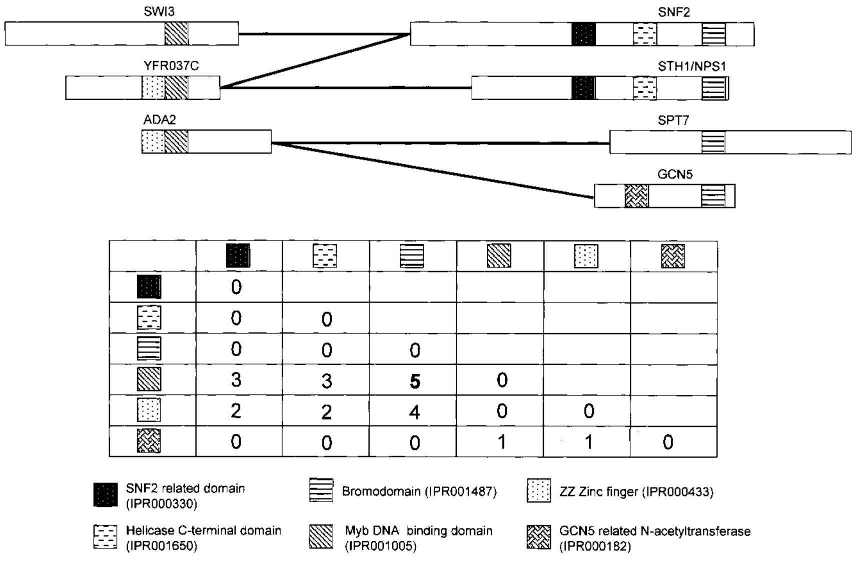
[a] The signature accession number in the InterPro database.

[b] The signature name in the InterPro database.

proteins. Likewise, only the *Bromodomain* is shared by all proteins in the second group. The *Myb* domain was defined in the retroviral oncogene v-Myb and its cellular counterpart c-Myb, and has been shown to be involved in DNA-binding.[23] The *Bromodomain* is found in many DNA-binding proteins of mammals, invertebrates and yeasts.[24] Recently, it has been shown experimentally that the *Bromodomain* of human histone acetyltransferase p300 is required for effective acetylation of the c-Myb protein, in addition to the well known histone acetyltransferase (HAT) domain. This finding supports the linkage detected by our analysis between proteins that contain the *Myb* domain and proteins that contain the *Bromodomain*. The acetylation of the c-Myb protein by p300 is not in the *Myb* domain but in its C-terminal region, as shown both *in vitro* and *in vivo*.[25] Thus, the signatures of *Myb* and *Bromodomain* can be used as indicators of a putative interaction between two proteins containing them, but it is not clear whether the *Myb* domain is involved in the molecular recognition.

The pairs of interacting proteins in Figure 2 could be predicted neither by gene adjacency considerations[17] nor by the phylogenetic profile approach.[20] The pairs of proteins in interaction that contain the *Myb* domain and *Bromodomain* are not adjacent, but reside on different chromosomes. Since both domains are found only in eukaryotes, the number of available completed eukaryotic genomes did not allow for a comprehensive phylogenetic profiling of these domains. However, further support for the linkage between the *Myb* domain and the *Bromodomain* was obtained by identifying protein sequences in which these two signatures were fused. Four proteins from the *Arabidopsis thaliana* genome were found to contain in their sequence both the *Myb* domain and the *Bromodomain* (O22724, O48523, O64877, Q9S7A8). Although none of the protein pairs in Figure 2 could be detected by a mutual sequence similarity to these fused proteins, we believe that the presence of both signatures in one protein provides further support for their use as indicators of protein-protein interaction.

Two other examples are demonstrated in Figure 3 and are briefly described here. Figure 3(a) demonstrates the five pairs of interacting proteins that show the sequence-signature pair of the *Cap-Gly* domain (IPR000938) and *CLUS00002* (see Table 2). As shown in the Figure, the three proteins with the *CLUS00002* signatures are very similar all along their sequences, while the *Cap-Gly* domain is the only shared signature among the second group's proteins. The *Cap-Gly* domain is a con-


SWI3
YFR037C
ADA2
SNF2
STH1/NPS1
SPT7
GCN5

	■	▥	▤	▧	░	▩
■	0					
▥	0	0				
▤	0	0	0			
▧	3	3	5	0		
░	2	2	4	0	0	
▩	0	0	0	1	1	0

■ SNF2 related domain (IPR000330)
▤ Bromodomain (IPR001487)
░ ZZ Zinc finger (IPR000433)
▥ Helicase C-terminal domain (IPR001650)
▧ Myb DNA binding domain (IPR001005)
▩ GCN5 related N-acetyltransferase (IPR000182)

Figure 2. The *Myb* domain-*Bromodomain* sequence-signature pair. Shown are the different protein pairs that contain these signatures (connected by the thick black line and denoted by their names in Swissprot). The three protein sequences with the *Myb* domain are: SWI3 (accession number P32591), transcription regulatory protein; YFR037C (P32591), hypothetical 63.2 kDa protein; and ADA2 (Q02336), potential transcriptional adaptor. The four protein sequences with the *Bromodomain* are: SNF2 (P22082), transcription regulatory protein; STH1/NPS1 (P32597), nuclear protein; SPT7 (P35177), transcriptional activator; and GCN5 (Q03330), transcriptional activator. The sequence alignment is demonstrated schematically, and the signatures (according to the InterPro database) are denoted as boxes with different patterns (legends for the signatures appear at the bottom of the Figure). The table summarizes the counts of the different sequence-signature pairs that appeared in the five pairs of interacting proteins.

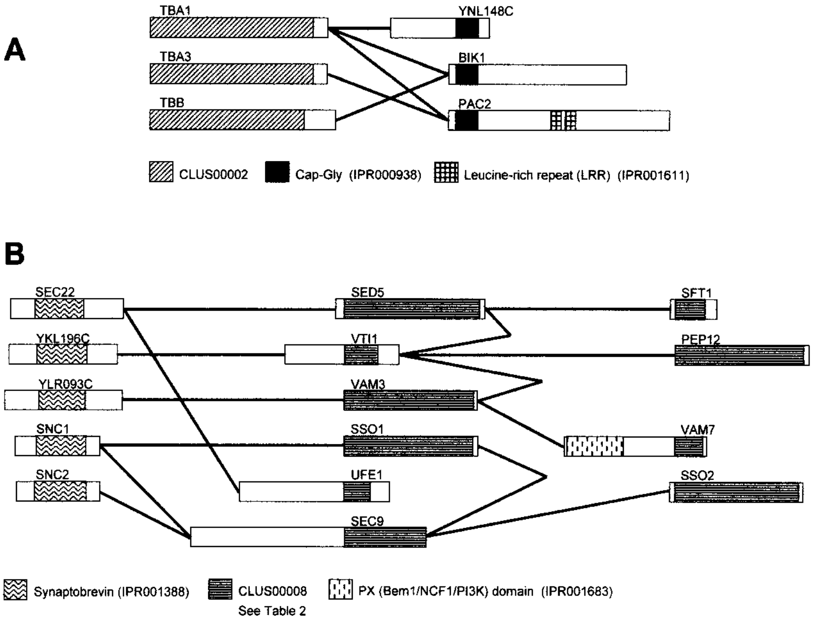


Figure 3. (a) The *Cap-Gly* domain, *CLUS00002* sequence-signature pair. The three protein sequences with the *CLUS00002* signature are TBA1 (P09733), tubulin α1 chain, TBA3 (P09734), tubulin α3 chain; and TBB (P02557) chain. The three protein sequences with the *Cap-Gly* domain are YNL148C (P53904), hypothetical 28.4 kDa; BIK1 (P11709), nuclear fusion protein; and PAC2 (P39937). For Figure details see the legend to Figure 2. (b) The *Synaptobrevin - CLUS00008* and *CLUS00008 - CLUS00008* sequence-signature pairs. The seven pairs to the left are those that show the *Synaptobrevin - CLUS00008* sequence-signature pair. The seven pairs to the right are those that show the *CLUS00008 - CLUS00008* sequence-signature pair. The five protein sequences with the *Synaptobrevin* domain are SEC22 (P22214), transport protein; YKL196C (P36015), hypothetical protein 22.7 kDa; YLR093C (Q12255), hypothetical protein; SNC1 (P31109), synaptobrevin homolog 1; and SNC2 (P33328), synaptobrevin homolog 2. The ten protein sequences with the *CLUS00008* domain, are PEP12 (P32854), syntaxin; SED5 (Q01590), integral membrane protein; VTI1 (Q04338), vesicle transport V-SNARE protein; VAM3 (Q12241), syntaxin; SS01 (P32867); UFE1 (P41834); SEC9 (P40357), protein transport; SFT1 (P43682); VAM7 (P32912), vacuolar morphogenesis protein; and SSO2 (P39926). For Figure details see the legend to Figure 2.

served glycine-rich signature that appears in many cytoskeleton associated proteins (CAP).[26] The signatures that were clustered in *CLUS00002* all belong to the tubulin family signatures (see Table 2). Members of this family are involved in polymer formation and are part of the microtubules.[27,28] Based on Watanabe *et al.*[29] the *Cap-Gly* domain is thought to be essential for association with microtubules. Thus, in this case the domain that was identified as the indicator of the interaction is the one that also plays a role in the molecular recognition.

As can be seen in Table 1, in several cases more complex interactions could be inferred by concatenation (e.g. if a-b is a signature combination and a-c is a signature combination, then proteins with signatures a,b, and c can be inferred to form a complex or be involved in a net of interactions, given that the signatures b and c are not derived from the same protein). Examples for putative higher-order interactions based on concatenation of correlated signatures are depicted in Figure 3(b) and include the signatures of *Synaptobrevin* (IPR001388) and *CLUS00008* (see Table 2). The proteins that include the signatures of *CLUS00008* interact both with other proteins with this signature and with proteins that include the *Synaptobrevin* signature. *CLUS00008* was defined by merging the *Syntaxin/epimorphin* family domain, *t-SNARE* coiled-coil domain, and *SNAP-25* family domain to one cluster (see Table 2). The *Synaptobrevin* family domain and the domains merged in *CLUS00008* are conserved from yeast to man, and proteins that contain them function in targeting and fusion of vesicles.[30] The identification of the *Synaptobrevin* (IPR001388) and *CLUS00008* signatures as indicators of complex formation is supported by experimental results showing that the rat SNAP-25 protein (exhibiting the *CLUS00008* signature) binds directly to both syntaxin protein (that contains also the *CLUS00008* signature) and synaptobrevin protein (containing the *Synaptobrevin* signature).[31]

The result of the concatenation procedure is a network of signatures, representing proteins that may form a complex or be involved in the same pathway. Thus, the concatenation of several correlated sequence-signatures defines a generic scheme for higher-order interactions, and different combinations of individual proteins that show these signatures may be predicted to be involved in these interactions, given that the concatenated signatures do not reside on the same protein. Several such concatenated relationships could be identified from the sequence-signature pairs listed in Table 1, and one such predicted generic high-order interaction is demonstrated in Figure 4.

Interestingly, there were quite a few over-represented sequence-signature pairs identified on the diagonal of the contingency table. These pairs are those that show the same sequence-signature in the two interacting proteins (Table 1). Such interactions may involve the same proteins or different proteins. One such example is demonstrated in Figure 3(b), where proteins with the *CLUS00008* signature interact with other proteins with this signature. Another example is provided by the sequence-signature pair IPR001163-IPR001163 (Table 1) that was derived from complexes of SM-like proteins. These proteins are known to be involved in forming the snRNPs and in binding to snRNA. None of these complexes in our data was of a homo-dimer, but all involved different proteins with the same sequence-signature. Thus, over-represented signature pairs that are identified along the diagonal of the contingency table can be used as indicators of interactions between the proteins that contain them, either for formation of homo or hetero-dimers, or for transient interactions that involve these proteins.

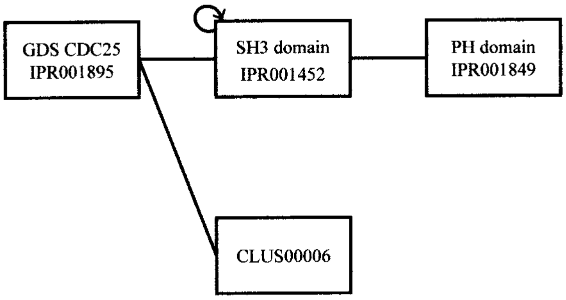


Figure 4. A network of signatures derived by concatenation of correlated signatures. Depicted in the Figure are the signatures that typify proteins that may be involved in higher-order interactions. For the signature names that compose *CLUS00006* see Table 2. The descriptions of the other sequence-signatures are according to the InterPro database.[22] IPR001895, guanine-nucleotide dissociation stimulators CDC25 family signature. IPR001452, Src homology 3 (SH3) domain. IPR001849, pleckstrin homology (PH) domain.

Use of the correlated sequence-signatures for prediction

We propose that the correlated sequence-signatures can be used as predictors of interaction. Namely, two proteins, where one contains the one signature and the other one contains the other signature, might interact. In order to evaluate the predictive power of the correlated sequence-signatures we chose to perform a cross-validation by a "leave-one-out" approach, due to the small size of the database. By this approach we exclude one protein pair at a time from the data set, re-construct the contingency table, and repeat the log-odds computation. We predict the excluded protein pair as interacting if the log-odds value of at least one of its sequence-signature pairs is above a certain threshold. This enables us to obtain an estimate of the sensitivity of the predictive approach (the number of correct predictions out of all known interactions). We first carried out this cross-validation test for interacting proteins with the sequence-signature pairs listed in Table 1, using the log-odds value of 2 as a threshold. We found that 94% of these protein pairs would have been predicted to interact, were they excluded from the learning set. This indicates that the sensitivity of the predictive scheme defined by the correlated sequence-signatures is very high. We repeated the sensitivity evaluation on another subset of the data, based on sequence-signature pairs with log-odds value of at least 1 and count of at least 3. Using the log-odds value of 1 as a threshold, we found that the sensitivity for this subset was 97%.

The specificity of the predictive approach can in principle be estimated by the fraction of correct predictions out of all predicted interactions in yeast. The latter can be computed based on the number of yeast proteins with the relevant signatures (two last columns in Table 1). However, until all of these interactions are tested experimentally and validated or rejected, it would be hard to estimate the specificity. Still, it is clear that the fraction of true positives out of the predicted interactions by the correlated sequence-signatures will be much higher than the fraction of these interactions out of all possible interactions in yeast. Thus, the usefulness of this approach is in increasing the specificity of the search for interacting proteins, in comparison to that expected by an unguided search in the space of yeast protein pairs.

The flowchart of the analysis that is demonstrated in Figure 1 depicts the most informative situation, where each protein is composed of several characteristic sequence-signatures, and there is no similarity between the proteins sharing a signature except in the signature region. There may be cases where two sequence-signatures are detected as correlated because very similar sequences were clustered within each of the two groups of interacting proteins. While it seems that in these cases there is no gain of information beyond simple sequence similarity, we believe that our clustering provides further support to the relationship between the two groups of sequences, especially when the similar protein sequences of one group interact with various, although similar in sequence, proteins of the other group. Nevertheless, we assessed the occurrence of such cases in our data by sequence comparison within each group of sequences corresponding to the sequence-signatures in Table 1, and found that for only three out of the 40 sequence-signature pairs in the Table were both sequence groups composed entirely of homologous sequences (IPR002161-IPR001852, IPR002161-IPR003009 and IPR001852-IPR003009).

Discussion

Deciphering the relationships between proteins in the cell is one of the major challenges of the post-genomic era. Since the interaction space that is defined by proteins encoded by a genome is enormous, experimental exploration of the interactions at a genomic scale may be overwhelming. An alternative approach would be to perform these screens in a directed manner, based on computational predictions. One appealing approach would be to predict the interacting partners by characteristic sequence motifs that typify the proteins that are involved in the interaction. Recently, Newman *et al.*[32] demonstrated the usefulness of this approach by concentrating on yeast proteins that are predicted to contain the coiled-coil domain. This structural domain is known to be responsible for molecular recognition, forming homo and hetero-dimers and higher-order multimers. When testing this group of proteins in yeast by the two-hybrid system, they succeeded in discovering many interactions that were not identified previously in the large-scale studies.[32] An apparent extension of this work should focus on other sequence motifs that may typify interacting proteins.

We propose that such characterization of the interacting proteins can be learned from the data accumulated in databases of experimentally determined interacting proteins. Such databases contain valuable information that can be processed and used in various ways. In several recent studies such data in *S. cerevisiae* were used to identify a set of links between proteins, defining a large protein network.[33-35] In the current study it is demonstrated that the information of experimentally determined interacting proteins could be used to classify sub-sets of protein pairs by some common denominator that can be subsequently used for prediction. We characterized pairs of interacting proteins by their sequence-signatures and showed that over-represented pairs of signatures could be identified. Some of the identified sequence-signatures coincide with domains that are known to be involved in the physical interactions (for example, the *Cap-Gly* domain). However, for many of the protein pairs in our database there is no knowledge about the domains involved in the interaction. Thus, the biological significance of the identified signatures with regard to the molecular recognition is not clear yet. The correlated sequence-signatures are presented here as markers that may typify proteins in interaction (like genetic markers are used). In a sense we propose to provide a classification of pairs of interacting proteins similar to available classifications of single protein sequences. A sequence-signature common to a group of sequences is used to characterize a protein family, and can be used to predict new putative family members. Similarly, in the current study we propose that correlated sequence-signatures can be used for characterizing two groups of sequences whose members are potential interacting counterparts.

The use of the correlated signatures as indicators of possible linkage between the proteins that contain them, although not specifying which proteins from the two groups interact, pinpoints a relatively small number of interacting candidates for further study. Using the correlated signatures for prediction dissects the huge space of all possible interactions of proteins encoded in a genome into smaller sub-spaces, each containing two groups of proteins that are predicted to interact. It is conceivable that carrying out directed interaction screenings in these sub-spaces is much more efficient than screening all possible interactions in a genome.[36] The number of tests that need to be made is much smaller and the probability of identifying interacting counterparts between the defined groups is higher. For example, in the yeast genome that encodes about 6000 proteins, there are theoretically about 1.8×10^7 possible interactions to test ($\sim 6000 \times 6001/2$). Suppose that as many as 1000 informative sequence-signature pairs are identified, and each of them defines two groups with 20 protein sequences on average in each. Then, the number of screens that need to be done is not greater than 4×10^5 ($1000 \times 20 \times 20$). Figure 5 demonstrates the predictions for the *Bromodomain* and the *Myb* domain that were discussed in detail in Results. There are 10 and 19 protein sequences that contain the *Bromodomain* and the *Myb* domain encoded in the yeast genome, respectively (note that there are many unknown proteins in the group of proteins that contain the *Myb* domain). Therefore, in this case 190 interactions are predicted, out of which 185 need to be tested experimentally (as five interactions were already included in the learning set). Such predictions can be applied to any genome of interest, based on the identified correlated signatures. In *C. elegans*, for example, there are 30 proteins with the *Myb* domain and 16 with the *Bromodomain*. We suggest that proteins from these two groups are likely to interact. Thus, the use of the correlated signatures in prediction of new interacting protein pairs in a genome reduces significantly the search space of putative pairs and increases the probability to detect such pairs.

As stated above, the use of the correlated signatures may narrow down significantly the number of candidates that are predicted to interact with a given protein, but cannot specify them exactly. Additional predictive parameters applied to the short list of candidates may aid in further pinpointing the actual interacting proteins. These include: phylogenetic profiling,[20] fusion analysis,[18,19] gene adjacency,[17] common cellular localization,[37] analysis of common regulatory sequences in upstream regions of the genes that encode the interacting proteins,[38,39] and analysis of correlated substitutions to identify proteins in interaction.[40] Also, with more refined classifications of the proteins it is possible that common multiple signature combinations in addition to the pairs will be identified. The additional signature combinations per pair of proteins will make the prediction more restrictive.

The identification of correlated sequence-signatures depends to a great extent on the quality and size of the experimental data. If the experimental data are incomplete (false negatives), some potentially informative signature pairs may be missed. The data may also include erroneous interacting proteins (false positives), as has been documented especially for the two-hybrid system.[41] These include sticky proteins, interactions identified erroneously due to auto-activation of the reporter genes, and interactions that are identified by the two-hybrid system but actually are not known to occur in the cell. The latter type of false positive is less problematic for the analysis we perform. Such false positives actually add information, as we are searching for the sequence-signatures that may typify the interaction, and for this purpose the fact that these proteins can interact is informative. Thus, when such pairs are clustered with true positive pairs they add to the possible identification of protein domains that either play a role in the interaction, or at least are indicative of the interaction.

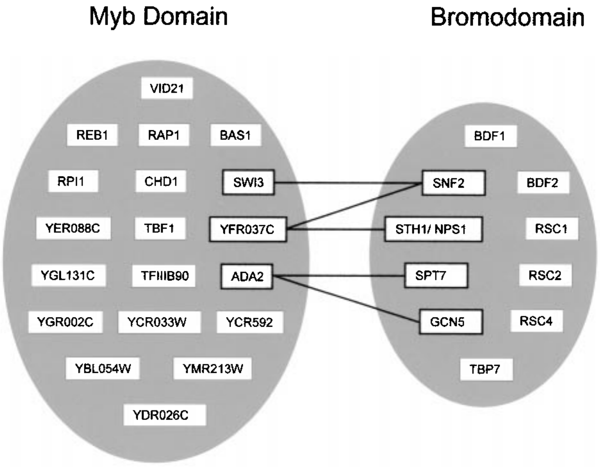


Figure 5. Prediction of putative interacting proteins that contain the signatures of the *Myb* domain and the *Bromodomain* in the yeast *S. cerevisiae*. There are 190 possible interactions between the proteins in the two groups: 185 interactions are new predictions. The five known interactions (included in the learning set) are marked by black lines.

The classification of the proteins by their sequence-signatures was found to be a serious limitation, as the data of yeast interacting proteins was reduced by 50 % due to lack of classification by the InterPro database for many of the protein pair members. A possible extension of the analysis may involve a search for possible new common motifs among the interacting proteins. Also, use of structural domains, in addition, or instead of the sequence-signatures, can be considered. Indeed, a classification of interacting proteins by their structural domains was recently reported.[42]

The small size of the data caused the contingency table to be very sparse. Despite the sparseness of the current database, the signature-signature table was found to be informative, as indicated by its information content of 2.48 bits. Thus, given a protein that is characterized by one of the InterPro signatures in the table, the information in the contingency table reduces the uncertainty involved in identifying its interacting partners. It is possible that even with a larger data set many of the empty entries will remain empty, because there are no interacting protein pairs that contain certain signature combinations. However, it is expected that with a larger data set the counts of the identified sequence-signature pairs in the non-zero entries will become higher and more significant. This refers especially to the entries with one observation. Extension of the database may be achieved by including data of interacting proteins from various organisms. However, for the type of analysis we perform, special attention needs to be paid to the sequence similarity between orthologs, to prevent a bias in the identified sequence-signature pairs.

We present a new classification scheme of interacting proteins whose basic unit consists of pairs of sequence-signatures. In practice the analysis generates a database of sequence-signature pairs that can be used for prediction: two proteins that exhibit these signatures are candidates for interaction. Another use of the correlated signatures may be in further support of other prediction schemes: identification of the defined correlated signatures in two proteins that are suspected to interact may be used to support the putative interaction. We regard the present analysis of the yeast data as a test case that exemplifies the potential of this approach. When more data of interacting proteins have been accumulated, and more extended classification of the proteins is available by the InterPro database or other sources, we expect better characterization of the correlated sequence-signatures. Our results suggest that this approach is valid for characterizing interacting proteins, and encourage the use of the correlated signatures as indicators of potential interacting proteins and for guided experimental interaction screens.

Methods

Experimental protein-protein interaction data

The protein-protein interaction data that were used in this analysis were collected from three different sources. (1) The Munich Information Center for Protein Sequences, Germany: MIPS.[43] This is a database of genomic data that includes information about various organisms. The yeast database MYGD, based on the genome of *S. cerevisiae*, contains a wealth of information about the yeast genome, including information on interacting proteins. It contains two tables of interacting proteins that differ by the experimental method used to identify the interacting proteins: the physical interaction table (~55 % of the interactions) and the genetic interaction table (~45 % of the interactions). The tables were downloaded on May 2000. (2) Data based on large-scale two-hybrid analysis in *S. cerevisiae*.[9] (3) The Database of Interacting Proteins: DIP.[35,44] This database documents experimentally determined pairs of interacting proteins from many organisms. For the current analysis only yeast proteins were used. The DIP flat file used in this analysis was dip030800.dat

We used these three sources of information to generate a non-redundant database that contained all documented interacting protein pairs in yeast. A protein that was included in more than one of the source databases was included only once in our database. Paralogs were maintained, as they may interact with different proteins and add to the identification of the correlated sequence-signatures. The total number of pairs in the database used in this study was 2908 (http://bioinfo.md.huji.ac.il/marg/Int-signature).

Protein classification by sequence-signatures

The InterPro database was used to describe the sequences by their characterizing signatures.[22] It is a new integrated documentation resource for protein families, domains and functional sites, developed as a non-redundant unified database which includes the classifications of PROSITE,[45] PRINTS,[46] Pfam[47] and ProDom.[48] The InterPro entries represent families, domains, repeats and sites of post-translational modification, that are encoded by regular expressions, profiles, fingerprints and hidden Markov models. The InterPro database includes sequence-signature characterization for all SWISSPROT and TrEMBL sequences. In the current analysis the InterPro data files from June 2000 were used. The data included 3052 different InterPro signatures. All the signatures of post-translational modifications were excluded. All yeast proteins were characterized based on the InterPro sequence-signature list. The InterPro database provides evaluations of the signature assignments along a protein sequence and flags the signatures accordingly. Only hits flagged as T, True and N, False Negative (when a sequence is known to belong to a certain family but the sequence-signature cannot be identified) were taken into account (see the InterPro User Manual for definitions).

Clustering of InterPro signatures

A parent/child relationship between two signatures is defined in the InterPro database when one signature entry describes a super-family and the other signature entry describes one of its sub-families.[22] The parent is the entry containing a more general signature of the family, while the children are more specific to certain members of the family. Therefore, the parent entry should contain more protein matches than each child. For example, for the tubulins there is a general pattern describing the whole family, and more specific patterns for the alpha tubulin, beta tubulin, etc. In order to further reduce the redundancy in signature definitions, clustering was performed based on the parent/child relationships. For example, in the tubulin case we treated all tubulin-specific signatures as one cluster of signatures (*CLUS00002)*.

A list of 253 different parent signatures and their children signatures was retrieved from the InterPro database. Based on this list, clustering was performed, resulting in 185 signature clusters. Based on the parent/child clustering, the characterization of all yeast protein sequences was repeated and, when relevant, the cluster number, instead of the original signature, was assigned to the sequence. The list of yeast sequence-signature clusters can be found at http://bioinfo.md.huji.ac.il/marg/Int-signature.

Acknowledgments

We are grateful to Yael Altuvia, Ora Schueler-Furman, Ruti Hershberg, Gila Lithwick and Esti Yeger-Lotem for their helpful comments on the manuscript. We also thank Yael Altuvia for the very efficient technical aid that she provided. Gill Bejerano and Naftali Tishby are thanked for helpful discussions, and Joel Sussman is thanked for providing computing facilities to E.S. while in Rehovot. The InterPro team is greatly acknowledged for their help. This study was supported by the Israeli Science Foundation administered by the Israeli Academy of Sciences and by The Israeli Ministry of Science.

References

1. Phizicky, E. M. & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**, 94-123.
2. Mendelsohn, A. R. & Brent, R. (1999). Protein interaction methods – toward an endgame. *Science,* **284**, 1948-1950.
3. Pandey, A. & Mann, M. (2000). Proteomics to study genes and genomes. *Nature,* **405**, 837-846.
4. Walter, G., Bussow, K., Cahill, D., Lueking, A. & Lehrach, H. (2000). Protein arrays for gene expression and molecular interaction screening. *Curr. Opin. Microbiol.* **3**, 298-302.
5. Nierman, W. C., Eisen, J. A., Fleischmann, R. D. & Fraser, C. M. (2000). Genome data: what do we learn? *Curr. Opin. Struct. Biol.* **10**, 343-348.
6. Lockhart, D. J. & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature,* **405**, 827-836.
7. Uetz, P. & Hughes, R. E. (2000). Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.* **3**, 303-308.
8. Cagney, G., Uetz, P. & Fields, S. (2000). High-throughput screening for protein-protein interactions using two-hybrid assay. *Methods Enzymol.* **328**, 3-14.
9. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al.* (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature,* **403**, 623-627.
10. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T. & Nishizawa, M., *et al.* (2000). Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA,* **97**, 1143-1147.
11. McCraith, S., Holtzman, T., Moss, B. & Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl Acad. Sci. USA,* **97**, 4879-4884.
12. Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P. *et al.* (2000). A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene,* **242**, 369-379.
13. Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S. *et al.* (2001). The protein-protein interaction map of *Helicobacter pylori*. *Nature,* **409**, 211-215.
14. Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A. *et al.* (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science,* **287**, 116-122.
15. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature,* **405**, 823-826.
16. Marcotte, E. M. (2000). Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* **10**, 359-365.
17. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324-328.
18. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science,* **285**, 751-753.
19. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402**, 86-90.
20. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA,* **96**, 4285-4288.
21. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature,* **402**, 83-86.
22. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E. & Biswas, M., *et al.* (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucl. Acids Res.* **29**, 37-40.
23. Biedenkapp, H., Borgmeyer, U., Sippel, A. E. & Klempnauer, K. H. (1988). Viral myb oncogene encodes a sequence-specific DNA-binding activity. *Nature,* **335**, 835-837.
24. Haynes, S. R., Dollard, C., Winston, F., Beck, S., Trowsdale, J. & Dawid, I. B. (1992). The bromodomain: a conserved sequence found in human, *Drosophila* and yeast proteins. *Nucl. Acids Res.* **20**, 2603.
25. Tomita, A., Towatari, M., Tsuzuki, S., Hayakawa, F., Kosugi, H., Tamai, K. *et al.* (2000). c-Myb acetylation at the carboxyl-terminal conserved domain by transcriptional co-activator p300. *Oncogene,* **19**, 444-451.
26. Riehemann, K. & Sorg, C. (1993). Sequence homologies between four cytoskeleton-associated proteins. *Trends Biochem. Sci.* **18**, 82-83.
27. Cleveland, D. W. & Sullivan, K. F. (1985). Molecular biology and genetics of tubulin. *Annu. Rev. Biochem.* **54**, 331-365.
28. Joshi, H. C. & Cleveland, D. W. (1990). Diversity among tubulin subunits: toward what functional end? *Cell Motil. Cytoskel.* **16**, 159-163.
29. Watanabe, T. K., Shimizu, F., Nagata, M., Kawai, A., Fujiwara, T., Nakamura, Y. *et al.* (1996). Cloning, expression, and mapping of CKAPI, which encodes a putative cytoskeleton-associated protein containing a CAP-GLY domain. *Cytogenet. Cell Genet.* **72**, 208-211.
30. Ferro-Novick, S. & Jahn, R. (1994). Vesicle fusion from yeast to man. *Nature,* **370**, 191-193.
31. Chapman, E. R., An, S., Barton, N. & Jahn, R. (1994). SNAP-25, a t-SNARE which binds to both syntaxin and synaptobrevin *via* domains that may form coiled coils. *J. Biol. Chem.* **269**, 27427-27432.
32. Newman, J. R., Wolf, E. & Kim, P. S. (2000). A computationally directed screen identifying interacting

coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **97**, 13203-13208.

33. Schwikowski, B., Uetz, P. & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnol.* **18**, 1257-1261.
34. Fellenberg, M., Albermann, K., Zollner, A., Mewes, H. W. & Hani, J. (2000). Integrative analysis of protein interaction data. *ISMB*, **8**, 152-161.
35. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M. & Eisenberg, D. (2001). DIP: the Database of Interacting Proteins: 2001 update. *Nucl. Acids Res.* **29**, 239-241.
36. Hu, J. C. (2000). A guided tour in protein interaction space: coiled coils from the yeast proteome. *Proc. Natl Acad. Sci. USA*, **97**, 12935-12936.
37. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005-10016.
38. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939-945.
39. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205-1214.
40. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.
41. Legrain, P., Wojcik, J. & Gauthier, J. (2001). Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet.* **17**, 346-352.
42. Park, J., Lappe, M. & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* **307**, 929-938.
43. Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A. *et al*. (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**, 37-40.
44. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucl. Acids Res.* **28**, 289-291.
45. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids Res.* **27**, 215-219.
46. Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P. *et al*. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucl. Acids Res.* **28**, 225-227.
47. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
48. Corpet, F., Servant, F., Gouzy, J. & Kahn, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucl. Acids Res.* **28**, 267-269.

Edited by G. von Heijne

(Received 7 March 2001; received in revised form 10 July 2001; accepted 10 July 2001)