**BMC Bioinformatics**

BioMed Central



# Implications for domain fusion protein-protein interactions based on structural information

*BMC Bioinformatics* 2004, 5:161 doi:10.1186/1471-2105-5-161

Jer-ming Chia (g04030676@nus.edu.sg)
Prasanna R Kolatkar (kolatkarp@gis.a-star.edu.sg)

**ISSN** 1471-2105

**Article type** Research article

**Submission date** 29 Jun 2004

**Acceptance date** 26 Oct 2004

**Publication date** 26 Oct 2004

**Article URL** http://www.biomedcentral.com/1471-2105/5/161

# Implications for domain fusion protein-protein interactions based on structural information

Jer-Ming Chia and Prasanna R.Kolatkar[§]

The Genome Institute of Singapore, Singapore

[§]Corresponding author

Email addresses:
JMC: g04030676@nus.edu.sg
PRK: kolatkarp@gis.a-star.edu.sg

# Abstract

## Background

Several *in silico* methods exist that were developed to predict protein interactions from the copious amount of genomic and proteomic data. One of these methods is Domain Fusion, which has proven to be effective in predicting functional links between proteins.

## Results

Analyzing the structures of multi-domain single-chain peptides, we found that domain pairs located less than 30 residues apart on a chain are almost certain to share a physical interface. The majority of these interactions are also conserved across separate chains. We make use of this observation to improve domain fusion based protein interaction predictions, and demonstrate this by implementing it on a set of *Saccharomyces cerevisiae* proteins.

## Conclusion

We show that existing structural data supports the domain fusion hypothesis. Empirical information from structural data also enables us to refine and assess domain fusion based protein interaction predictions. These interactions can then be integrated with downstream biochemical and genetic assays to generate more reliable protein interaction data sets.

## Background

Networks of interacting molecules drive every process in biological cells. Proteins dominate these networks, some of which involve transient interactions such as signal transduction cascades and ligand-receptor interactions, while others form more permanent molecular machineries such as ribosomes and polymerases. Unraveling these networks and interactions will not only help us better understand complex cellular processes, but also enable us to make inferences about the function of individual proteins through 'guilt-by-association' [1].

Over the last few years, high-throughput interaction detection assays have been introduced and refined to complement the traditional genetic and biochemical techniques. High-throughput mass spectrometry protein complex identification (reviewed by Pandey and Mann [2]), and yeast two-hybrid systems [3] are examples of these. The success of these techniques is well illustrated in the budding yeast *Saccharomyces cerevisiae*, in which networks of its interacting proteome where constructed using genome-wide screens. [4-7]

The wealth of genomic and protein sequences, the increase of 3D structures of protein complexes, together with the deluge of microarray expression data, has provided researchers with an overwhelming body of information that can be used to infer both functional as well as interaction linkages. Clearly, bioinformatics and computational biology are necessary tools for delineating this information. In response to the data explosion, several *in silico* methods have recently been developed to predict associations from these data.

Phylogenetic profiles focus on the co-occurrences of genes across several organisms. By studying the pattern of evolutionary conservation between sets of genes in different organisms (phylogenetic distribution), it has been shown that these phylogenetic profiles can be successfully used to infer both localization as well as functional association between proteins [8-10]. Protein domains that are found fused together within a protein are frequently involved in the same process, and in many examples proven to be physically interacting. This phenomenon is the basis for the domain fusion analysis, which can be used to predict protein interactions in cases where the fused domain pair is found independently across separate protein chains [11, 12].

Structural data has also been mined and analyzed for residue patterns within interfaces between pairs of interacting proteins. These are then used to train learning models for *ab initio* categorization and prediction of protein interactions [13, 14]. Jansen and co-workers [15] illustrated how expression profiles from mRNA expression data could be harnessed and used as an effective source for the prediction of protein interactions.

A number of groups have compared and reported on the protein interaction datasets that are emerging from the various genome-scale biochemical, genetic and *in silico* experiments [16-18]. All of them drew a similar conclusion; high-throughput methods produce little overlapping results, and taken singularly, each technique has a high error

rate (false positive and false negative). Each of these methods has their own specific strength and weakness, and covers a separate subset of interactions. Integrating the various result sets together, allows one to piece together a map of the interacting proteome that is more reliable with higher accuracy, and more informative with higher coverage.

The study by von Mering and co-workers [17] showed that *in silico* methods have higher coverage and higher accuracy than the majority of biochemical/genetics methods, second only to high-throughput mass spectrometry. The use of sensible strategies and filters has allowed *in silico* analyses to provide better performance. On top of that, these methods are less biased towards abundant proteins. *In silico* analyses are indispensable, and further improvements of these methods to make them more accurate will provide a cleaner set of data for downstream biochemical/genetic studies.

In this study, we make use of an empirical observation that domain pairs, which lie in close proximity on a protein chain tend to interact, to refine the domain fusion analysis. This way, we aim to improve the accuracy of the domain fusion analysis.

### Domain fusion

The basis for domain fusion (or gene fusion) is the observation that certain proteins (termed the Rosetta stones) in a given species are found to consist of a fusion between two separate proteins in another species. Through fusion, the entropy of dissociation between the two proteins is reduced, and it is hypothesized that in all likelihood, these two separate proteins share a functional association, if not a physical interaction [11, 12].

Domains have been described as the primary building blocks of proteins [19], recombining in various permutations, resulting in proteins of completely different functions [20]. In our implementation of the domain fusion analysis, we chose the representation of proteins being composed of domains, separated by linkers on a peptide chain.

In this paper, we make use of existing structural data to support the domain fusion hypothesis. We interrogated known 3D structures for evidence of inter-domain physical interactions on the same chain. We investigated and concluded that there was an association between the distances at which domains are spaced apart on the chain, and the propensity for a domain-pair to interact.

We also show that domain pairs, located in close proximity on a protein chain, are likely to interact even when found residing on different chains, hence proving that the domain fusion hypothesis is valid.

Finally, we demonstrate that peptide chains with closely spaced domains are likely to make better Rosetta stones, and we make use of this observation to improve domain fusion based protein interaction predictions.

# Results

### Supporting the domain fusion hypothesis

The available structural data indicate that intra-chain domain pairs, which lie in close proximity on a peptide chain, tend to physically interact with one another. The mean distances of interacting intra-chain domain pairs are smaller than ones which do not interact; interacting pairs are on the average 50 residues apart, while non-interacting pairs have a mean distance of 166 residues between them.

In order to verify the correlation between distance and interaction, we made use of contingency tables and the chi-squared test statistic. For a set of inter-domain distances ranging from 5 residues to 200 residues, we constructed 2x2 contingency tables that classified domain pairs according to two criterions; 1) whether or not they are separated by a distance no greater than a threshold, and 2) whether or not the domain pair is interacting. The chi-squared value of each table was used as a statistic to test the null hypothesis ($H_0$): Domains pairs separated by no greater than a predefined distance and their tendency to interact were independent. The p-values indicate the probability of having the chi-squared test statistic as extreme as, or larger than observed when $H_o$ is true.

We found that the contingency table for domain pairs spaced up to 30 residues apart had the highest chi-squared value, with a statistically significant p-value of less than 0.001, allowing us to confidently reject $H_0$. This trend is noticeable in the chart illustrating the proportion of interacting pairs across various inter-pair distances (figure 1). Domain pairs located less than 30 residues apart are almost certainly (90%) to be in contact with each other, whereas only half (51%) of domain pairs with more than 30 residues separation were categorized as physically interacting. The chi-squared value is also overlaid on the chart in a dotted line, representing the test statistic from each corresponding contingency table.

In order to validate the domain fusion hypothesis, we not only need to show that domain pairs on the same chain tend to interact with each other, but importantly, this same domain pair will tend to be in contact if they are located independently across separate chains of a polypeptide complex. From our data, we noticed that 71% of domain pairs, which lie within 30 residues of each other on the same chain, could be found physically interacting across separate chains of a complex. In contrast only 38% of domain pairs lying greater than 30 residues apart are seen to be in contact within a multi-chain complex. Once again, putting this into a contingency table and evaluating the chi-squared statistic we reject the null hypothesis (p-value <= 0.001). In other words, there is a correlation between domain-pairs spaced less than 30 residues apart on a single peptide chain and their tendency to interact across separate chains of a polypeptide complex.

**30 residues criteria applied to Swiss-Prot proteins**

We wanted to verify that the 30 residues criteria could be used as a measure to filter and improve predictions made using the domain fusion methodology. A set of proteins for the budding yeast *S. cerevisiae* was downloaded from Swiss-Prot, and domain fusion based protein interactions were predicted as described in the Methods section. After filtering for promiscuous domains, a total of 9279 protein interactions remained, of which 28% or 2629 were supported by a Rosetta stone with no more than 30 residues between the fused domains.

The functional category assigned to each protein in an interacting pair was used to gauge the plausibility of the interaction; if two different proteins were found physically interacting, one would expect the two proteins to have overlapping functional categories. 62% of the interacting protein pairs, supported by a 30 residue Rosetta stone, have both partners belonging to the same functional category. The same proportion for interacting pairs not supported by a 30 residue Rosetta stone is 48%. This 14% difference is significant with a p-value of less than 0.001, using a two-sample t-test.

# Discussion

*In silico* methods for predicting protein interactions are not only able to match the accuracy of the other genetic, biochemical and biophysical techniques, but also have the added advantage of providing higher coverage [17]. Among the *in silico* methods, domain fusion is an attractive technique because it enables a functional link to be drawn between two proteins based solely on their primary sequence. Still, large-scale sets of high-throughput protein interaction data available today are spurious, more than half of them proving to be false positives [17], the challenge remains to improve the quality of high throughput protein interaction data sets.

Protein interactions can be classified as either permanent or transient interactions. The data from this study were taken from the PDB, where most of the submitted structures are results from x-ray crystallography experiments. Consequently, we believe that the vast majority of our deduced domain and protein interactions are physical, permanent interactions.

Our study of multi-domain, single and multiple-chain protein structures in the PDB gave us two results. First of all, it supports the domain fusion hypothesis suggested by Marcotte and Enright. Secondly, it allows us to conclude that single chain peptides with closely spaced domain pairs make better Rosetta stones, and hence better predictors of protein interactions.

Evident from the set of PDB structures we studied, is a correlation between the distance separating a pair of domains on a protein chain, and their tendency to physically interact with one another. As described by Marcotte and co-workers when they constructed the domain fusion hypothesis for evolution of protein interactions, affinity between interacting pairs of domains may be enhanced when the domains are fused together on the same chain [11]. Consequently, close proximity of the interacting pair on the same chain increases the effective local concentration of the two domains, facilitating the interaction. The biochemical advantage for such an arrangement would explain the tendency for interacting domains to be found close together on a protein chain. Our observation that domain pairs located less than 30 residues apart are almost certainly to share an interface clearly supports this idea.

Previously, Park and co-workers [21]had observed this figure in an unrelated report. In this study, we adopted a different concept of a protein domain - PFAM categories which are essentially sequence-based annotations. Analyzing a substantial set of structural data from the PDB, we also derive at this similar threshold of 30 residues, and show it to be statistically significant.

### Conservation of domain Interactions across multi-chain structures

The data from multi-chain PDB structures provide additional support to the domain fusion hypothesis, by showing that most of the intra-chain domain interactions are similarly represented across separate chains of a complex. This provides additional

mechanistic evidence that the interaction between the two domains is most probably functional and conserved.

To our knowledge, this is the first time structural data has been used to support the domain fusion hypothesis.

### Functional classification of non-interacting domains in close proximity

We tried to uncover a pattern within the set of closely spaced, yet non-interacting domain pairs. We wanted to detect if there was an over-representation of domains from a specific molecular functional category in this non-interacting list. This list is displayed in Table 1. From the Gene Ontology categories of the domains, it is obvious that a good proportion of domains on the list are involved in DNA/RNA processing activities, as well as catalytic functions, but we didn't observe any statistically significant differences when comparing this non-interacting set with the sets of domain pairs which interact. This could be due to the small number of non-interacting domains in close proximity.

Furthermore, since the interactions we can detect from structural data are more likely to be permanent interactions, it is possible that the reason no physical contact is witnessed between these proximal domains in structural data is because the domains form transient interactions that are not captured in the x-ray crystallography data.

### Hot loops and interactions

We also looked for a relation between protein disorder and interacting domain pairs. We wanted to see if protein domain pairs which interact on the same chain, tend to be linked by a disordered region. To this effect, we used DisEMBL[22]to do the disorder analysis. However, we were unable to infer any relationship between disorder and interacting domains.

### Use of 30 residue criteria to refine domain fusion predictions

Our results from predicting interactions among *S. cerevisiae* proteins indicate that Rosetta stones with domains separated by less than 30 residues do indeed make better domain interaction (and hence protein interaction) predictors.

The set of protein interactions inferred from these Rosetta stones are enriched with more reliable interactions, as judged by using similar function as a criteria. The total number of interactions is reduced to nearly a quarter when employing this method. This allows us to conclude that the number of false positives is reduced, increasing the accuracy of the prediction. Without needing to employ a hard filter, protein interactions predicted using the domain fusion methodology may be ranked according to the quality of the Rosetta stones each interaction is inferred from, allowing one to identify a much smaller subset of more reliable interactions, and use them for downstream analyses.

## Conclusions

We have successfully demonstrated the use of current structural data as a resource for refining current protein interaction predictions, in particular domain fusion predictions. Our data strongly suggests that domain pairs separated by less than 30 residues on a peptide chain are almost certainly to physically interact, and this criterion is useful in accessing protein interactions predicted from Rosetta stone proteins.

Going forward, the availability of a large number of structures through structural genomics programs will facilitate a larger sampling of the domain structure space. New patterns may emerge as use of this data becomes available, allowing better predictions to be made.

# Methods

### Intra-chain domain interactions

We used domain models from the Protein Family database (PFAM) [23] which were mapped onto structures from the Protein Databank (PDB) [24]. The PFAM to PDB mappings were obtained from PFAM data files, and we only considered PFAM entries that were tagged with the type 'Domain'. There are a total of 4169 peptide chains in the PDB that are annotated with more than one PFAM domain, comprising a total of 504 unique PFAM domains present within the data set. In order to obtain a non-redundant representation of these peptide chains, we took clusters of them based on 50% sequence identity, and selected one representative from each cluster. This left us with a set of 565 3D structures of multi-domain peptide chains, comprising a total of 996 distinct domain pairs, of which 478 are unique pairs.

We used the coordinates within the PDB data files to calculate the distances between domains, and to determine if they are interacting. Two domains are judged to be interacting if they share at least five contacting residue pairs, where contacting residues are residue pairs with less than 6Å between their respective alpha-carbon atoms.

### Multi-chain interactions

Using a similar approach to the above, we obtained a set of multi-chain PDB structures in which the previously determined domain-domain interactions can be observed across separate peptide chains within a complex. The ASALIST from the PQS server [25]was used to sift out the biologically significant contacts from the crystal packed structures. Of the 379 domain pairs above, 305 were found on separate chains of a complex, and these were used for the analysis.

### GO functional annotation

PFAM domains were categorized into Gene Ontology (GO) molecular functions and cellular processes using the PFAM2GO data provided by the GO consortium [26].

### *Saccharomyces cerevisiae* protein interactions prediction

In order to assess how distance between domains could be used to improve the domain fusion based protein interaction predictions, we predicted interactions between 6918 proteins from the organism *Saccharomyces cerevisiae* found within the Swiss-Prot database, and gauged the quality of the interactions by looking at the function of each interacting protein.

The steps taken to predict protein interactions based on domain fusion are as follows. Swiss-Prot (release 42.9) and Trembl (release 25.9) [27] protein datasets were first searched for multi-domain proteins, by relying on their PFAM annotations. As above, only PFAM domains of type 'Domain' where considered. These multi-domain proteins

were then catalogued as Rosetta Stones. Pairwise domain interactions were inferred by cataloging each distinct domain pair found on every Rosetta Stone protein, together with the number of residues separating the pair. As described by Marcotte and co-workers [11], domain interactions involving the 5% most promiscuous domains were discarded, removing the majority of false positives.

This domain interaction set was then used to predict pairwise protein interactions between the *S. cerevisiae* proteins, by looking at the complement of PFAM domains between each and every pair of proteins, and seeing if there were any Rosetta stone determined domain interactions between the domains of each protein. The protein interactions were sorted into two groups; one group inferred from domain interactions supported by the existence of a Rosetta stone protein with no more than 30 residues between the domain pair, and the other group with no support from a 30 residue Rosetta stone.

To validate these protein interactions, we mapped the proteins to the MIPS comprehensive yeast genome database [28], and looked for interacting protein partners that share the same MIPS functional category. Interactions between pairs that share the same function are more likely to be true.

All the data was stored in a relational database schema implemented in MySQL, with a set of perl modules written for data transaction and manipulation. The Bioperl bioinformatics tool kit [29] was used to parse Swiss-Prot, Pfam and PDB data, as well as to extract coordinates of each atom from each PDB structure.

## Authors' Contributions

JMC participated in the design of the study, performed the data and statistical analysis, as well as drafted the manuscript. PRK conceived of the study, participated in its design and coordination, and edited the final draft of the manuscript. Both authors read and approved the manuscript.

## Acknowledgements

This work is supported by the Agency for Science, Technology and Research (A*STAR) in Singapore. We would also like to thank Dr Radha Krishna Murthy and Dr Li Yi for their valuable input.

## References


1. Oliver S: **Guilt-by-association goes global**. *Nature* 2000, **403**(6770):601-603.
2. Pandey A, Mann M: **Proteomics to study genes and genomes**. *Nature* 2000, **405**(6788):837-846.
3. Fields S, Song O: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**(6230):245-246.
4. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**(6868):141-147.
5. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**(6868):180-183.
6. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.
7. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623-627.
8. Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D: **Localizing proteins in the cell from their phylogenetic profiles**. *Proc Natl Acad Sci U S A* 2000, **97**(22):12115-12120.
9. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci U S A* 1999, **96**(8):4285-4288.
10. Huynen MA, Bork P: **Measuring genome evolution**. *Proc Natl Acad Sci U S A* 1998, **95**(11):5849-5856.

11. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**(5428):751-753.
12. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature* 1999, **402**(6757):86-90.
13. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology**. *Proc Natl Acad Sci U S A* 2002, **99**(9):5896-5901.
14. Ofran Y, Rost B: **Analysing six types of protein-protein interfaces**. *J Mol Biol* 2003, **325**(2):377-387.
15. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **12**(1):37-46.
16. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes**. *Trends Genet* 2002, **18**(10):529-536.
17. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.
18. Bader GD, Hogue CW: **Analyzing yeast protein-protein interaction data obtained from different sources**. *Nat Biotechnol* 2002, **20**(10):991-997.
19. Copley RR, Doerks T, Letunic I, Bork P: **Protein domain analysis in the era of complete genomes**. *FEBS Lett* 2002, **513**(1):129-134.
20. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution**. *Nature* 2002, **420**(6912):218-223.
21. Park J, Lappe M, Teichmann SA: **Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast**. *J Mol Biol* 2001, **307**(3):929-938.
22. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure (Camb)* 2003, **11**(11):1453-1459.
23. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32 Database issue**:D138-141.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
25. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server**. *Trends Biochem Sci* 1998, **23**(9):358-361.
26. 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: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
27. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The**

**SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.

28. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 2002, **30**(1):31-34.
29. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**(10):1611-1618.

**Figure 1. Distance between domain pairs on a protein chain and the likelihood that they interact**

The solid line indicates the percentage of domain pairs, within a distance range apart, which are in contact. The broken line shows the distribution of chi-squared values corresponding to constructed 2X2 contingency tables that classified domain pairs according to 2 criterions; 1) whether or not they are separated by a distance no greater than the upper limit of each range, and 2) whether or not the domain pair is interacting.
The percentage of interacting domain pairs drop noticeably after 30 residues, and the chi-squared value is also maximum at this threshold.

**Table1: List of domain pairs separated by less than 30 residues but are not interacting**

| Domain 1 | Molecular Function | Domain 2 | Molecular Function | No. of Contacts |
|---|---|---|---|---|
| 2-Hacid_DH | oxidoreductase activity | 2-Hacid_DH_C | oxidoreductase activity | 4 |
| CH | | CH | | 0 |
| Cytochrome_CIII | | Cytochrome_CIII | | 0 |
| dsrm | double-stranded RNA binding | dsrm | double-stranded RNA binding | 0 |
| EGF | | EGF | | 0 |
| eRF1_1 | | eRF1_2 | | 0 |
| FKBP | | FKBP | | 1 |
| fn1 | | fn1 | | 0 |
| fn1 | | fn2 | | 2 |
| fn2 | | fn2 | | 0 |
| GlutR_NAD_bind | glutamyl-tRNA reductase activity | GlutR_dimer | glutamyl-tRNA reductase activity | 1 |
| HTH_9 | molybdate ion transporter activity | TOBE | | 0 |
| kazal | | kazal | | 0 |
| MHC_II_beta | immune response | ig | | 2 |
| myb_DNA-binding | DNA binding | myb_DNA-binding | DNA binding | 0 |
| Peptidase_M10 | proteolysis and peptidolysis | fn2 | | 2 |
| Phe_tRNA-synt_N | phenylalanine-tRNA ligase activity | tRNA-synt_2d | phenylalanine-tRNA ligase activity; | 0 |
| resolvase | recombinase activity;DNA recombination | HTH_7 | recombinase activity;DNA recombination | 1 |
| RHD | regulation of transcription, DNA-dependent | TIG | | 0 |
| Ribosomal_L9_N | structural constituent of ribosome | Ribosomal_L9_C | structural constituent of ribosome | 0 |
| RNase_PH_C | RNA binding;RNA processing | KH | nucleic acid binding | 0 |
| Rotamase | isomerase activity | Rotamase | isomerase activity | 0 |
| rrm | | rrm | | 4 |
| rve | DNA binding;DNA recombination | integrase | integrase activity | 0 |
| SH2 | intracellular signaling cascade | SH3 | | 0 |
| sushi | | sushi | | 4 |
| TPP_enzymes_N | | TPP_enzymes | | 2 |
| WW | | WW | | 0 |
| zf-Sec23_Sec24 | | Sec23_trunk | | 0 |

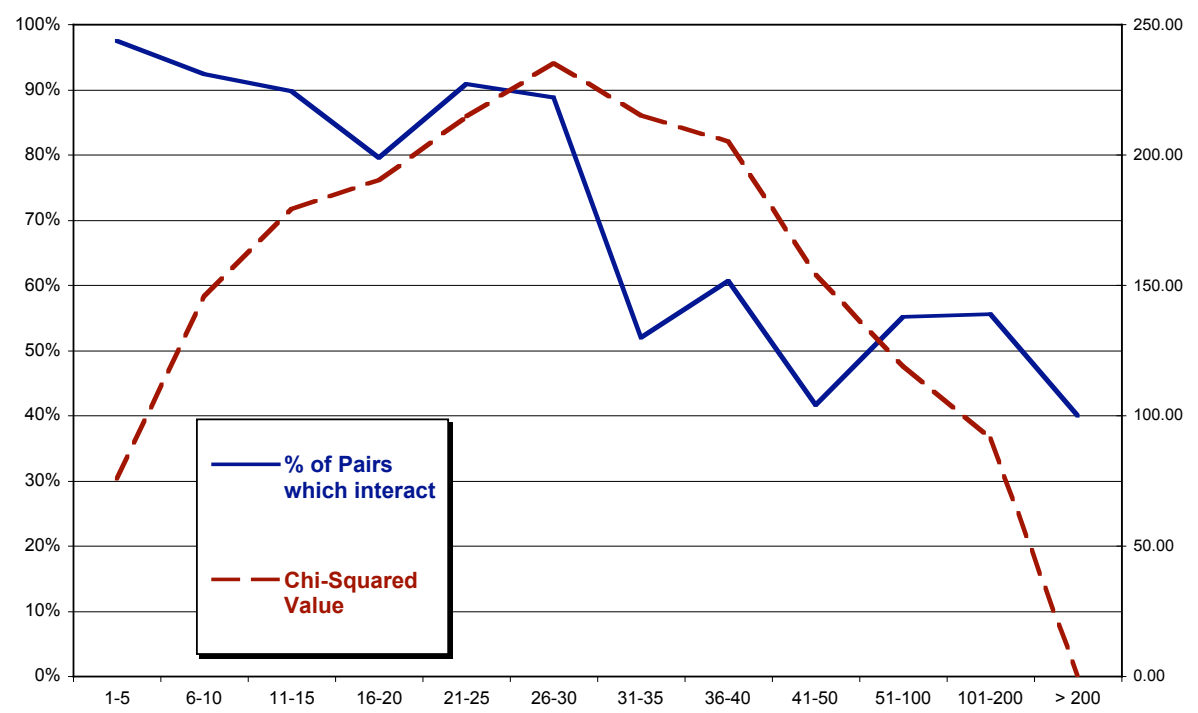
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
250.00
200.00
150.00
100.00
50.00
0.00
% of Pairs which interact
Chi-Squared Value
1-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-50
51-100
101-200
> 200
Distance between domains pairs

Figure 1