Using Reliable and Surprising Item Sets for the Characterization of Protein-Protein Interfaces

Christine Martin[12], Antoine Cornuéjols[2]

[1] LIMSI-CNRS, Université Paris-Sud, UPR CNRS 3251
Bâtiments 508 et 502bis 91403 ORSAY (France)
christine.martin@limsi.fr
[2] AgroParisTech, Equipe Statistique et génome, UMR AgroParisTech/INRA 518
Département M.M.I.P., 16 rue Claude Bernard 75005 Paris Cedex France
antoine.cornuejols@agroparistech.fr

Abstract: *Numerous research effort have been aimed to characterize and predict protein-protein interfaces. This paper introduces a method using only known protein-protein interfaces and combining frequent item set mining techniques with statistical tests to ensure the selection of interesting features. Starting from a database of known interfaces described with geometrical elements, the method produces the elements and combinations thereof that are* characteristic *of the interfaces. This approach allows one to eliminate the need for negative instances and to come up with easy to interpret features, as compared to techniques that operate as "black-boxes". The results obtained on a set of 459 protein-protein interfaces from the DOCKGROUND database confirm that the findings are consistent with current knowledge about protein-protein interfaces.*

Keywords: Protein-protein docking, data mining, frequent itemsets.

1 Introduction

Being able to predict protein-protein interactions is of paramount importance for biology and medicine [6,10]. A first approach is to *start from fundamental premisses*, i.e. the sequences of amino-acids that make up the proteins, and knowledge of their physico-chemical properties and how these translate in terms of energy. In principle, a sufficiently detailed model should allow one to compute with enough accuracy the energy of each configuration of interest and therefore predict the likely protein-protein complexes and their probable binding sites. This line of attack is however precluded, at least at the present time, by the sheer magnitude of the size of the search space and by the complexity of the energy computations. Another route is to *learn by automatic means to discriminate positive protein-protein complexes from negative ones* [3]. One issue is the choice of representation of the protein-protein complexes. Another one stems from the fact that, usually, databases only contain positive instances. Negative instances have to be generated, often using random conformations and orientations of the proteins, assuming that these correspond to bad or impossible pairings. However, this strategy may be disputed and can profoundly affect the performance of the learning methods.

This is why another approach is proposed here. In our work, interfaces of known protein-protein complexes are described by collections of small subgraphs taken in a dictionary of elementary patterns, much as transactions in a database of purchases in a supermarket are made of collections of items in a given set of products. This analogy suggests to use data mining techniques in order to detect characteristic regularities in known protein-protein interfaces.

After briefly describing the representation of the protein-protein interfaces, section 2 describes in a generic way the proposed method to analyze whether the interfaces of protein-protein complexes have special properties, and, if yes, which ones. The results obtained using data extracted from the Dockground database [7] are described in section 3. Finally, section 4 discusses the results obtained in light of the overall protein docking problem and opens directions for extensions of this work.

2 The aCID method for the characterization of protein-protein interfaces

In our work, interfaces of known protein-protein complexes are described by collections of *edges*, *triangles* and *tetrahedra* extracted from a geometric representation of proteins called weighted α-complexes [5,8] and completed by additional information about the nature of the amino acids involved in each of them. To obtain a reasonable number of descriptors as compared to the amount of available data, we grouped the amino acids with respect to their physico-chemical properties [4] : *Hydrophobic*, *Polar*, *positively charged*, *negatively charged* and *Small* (resp. noted H, P, $+$, $-$ and S). This leads to a repertoire of 120 distinct descriptive items.

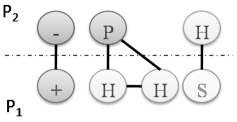

Figure 1. An example of a protein-protein interface in the chosen representation.

In our dataset (see section 3), one typical interface involves between 15 to 50 items, some of them possibly repeated (e.g. figure 1). On average, each interface contains 22 geometrical items, which gives rise to approximately 10,000 items for the whole set of the 459 interfaces studied.

The central question is: *do the interfaces of the known protein-protein complexes present special regularities*?

2.1 Analysis of the frequencies of items

Suppose we observe that a given item, say (SHP), occurs 150 times in all (over the 10,000 items taken altogether) and is present in 50 out of the 459 interfaces. What should we think? Is this feature normal? Mildly surprising? Quite astonishing? One that could be used as a "signature" of a likely interface between proteins? To answer these questions necessitates that expectation under "normal circumstances" be defined (a.k.a. as a "null hypothesis") and that deviations from it can give rise to probability assessments.

In order to compute the probability associated with each item $\mathtt{A}$ (e.g. $(S++)$), one can measure the probability that it would appear as the result of the combination of half-items $\mathtt{A}_i\mathtt{A}_j$ (e.g. $(S++)$ could result from $(S \leftrightsquigarrow ++)$ or from $(+ \leftrightsquigarrow S+)$). In general, given that an event $\mathtt{A}$ can result from pairs of sub-events $\mathtt{A}_i\mathtt{A}_j$, its expected number $n_{\mathtt{A}}$ under the binomial assumption[3] is:

$$\mathbb{E}[n_{\mathtt{A}}] = \sum_{i,j} a_{ij} \cdot p(\mathtt{A}_i) \cdot p(\mathtt{A}_j) \cdot N \tag{1}$$

[3] I.e. independent and identical trials.

where $p(X)$ is the probability of the item X as measured in the interfaces, N is the total number of events and $a_{ij} = 1$ if $A_i = A_j$ or $a_{ij} = 2$ if $A_i <> A_j$. And the variance is given by:

$$Var[n_{\mathtt{A}}] = \big(\sum_{i,j} a_{ij} \cdot p(\mathtt{A}_i) \cdot p(\mathtt{A}_j)\big)\big(1 - \sum_{i,j} a_{ij} \cdot p(\mathtt{A}_i) \cdot p(\mathtt{A}_j)\big)N \quad (2)$$

For instance, suppose again that one is interested in the $(S++)$ item. One would measure the probability of having the semi-item (S), $(+)$, $(++)$ and $(S+)$ which would enable to get: $\mathbb{E}[n_{(S++)}] = 2\big(p(S)\cdot p(++) + p(+)\cdot p(S+)\big)\ N$, where $p(x)$ would be the observed frequency of the semi-item x in all semi-interfaces[4], and N be the number of items in all 459 interfaces, that is 10,000 (the factor 2 comes from $\big(\sum_{i,j} p(\mathtt{A}_i)\cdot p(\mathtt{A}_j)\big) = \big(\sum_{i,i} p(\mathtt{A}_i)\cdot p(\mathtt{A}_i)\big) + 2\big(\sum_{i,j,i<j} p(\mathtt{A}_i)\cdot p(\mathtt{A}_j)\big)$ which reflects the fact that the same item can be obtained with an $\mathtt{A}_i$ coming from either semi-interface).

Note that no combination of semi-items should be considered that lead to items with more than 4 elements (tetrahedra). This must be taken care of in the computation of formula 1 and 2.

2.2 Analysis of the frequencies of the combinations of items

We look for combinations that would be very differently represented than what should be expected under a null hypothesis where the items would be independent. In general, the expected number of a m-combination of m items $\underbrace{\mathtt{A}_i, \mathtt{A}_j, \ldots, \mathtt{A}_k}_{m}$ is:

$$\mathbb{E}[n_{\mathtt{AB}}] = \Pi_{l=i,j,\ldots,k}\ \ a_{i,\ldots,k} \cdot p(\mathtt{A}_l) \cdot N$$

and the variance:

$$Var[n_{\mathtt{A}}] = \big(\Pi_{l=i,j,\ldots,k}\ \ a_{i,\ldots,k} \cdot p(\mathtt{A}_l)\big)\big(1 - \Pi_{l=i,j,\ldots,k}\ \ a_{i,\ldots,k} \cdot p(\mathtt{A}_l)\big) \cdot N$$

with N the number of observed items of size k and $a_{i,\ldots,k}$ the number of permutations of $A_i, \ldots, A_k$.

2.3 Combing the items and combinations

Underrepresented items are not to be retained if the goal is to discover elements that are responsible for the binding of protein-protein complexes. We keep therefore the items (or the combinations of items) of which the observed number in the known interfaces exceeds its expected number by more than twice the standard deviation: $n_X^{obs} \geq \mathbb{E}[n_X] + 2\sqrt{Var[n_X]}$. Under the normal distribution assumption, the probability of observing n_X^{obs} events or more is then less that 2.5% [5]. The choice of this threshold controls the rate of false positive elements (Type 1 error)[6].

In the same spirit, we would rather identify elements that seem well correlated to as large as possible a fraction of all known interfaces. This means both that they are significantly over represented (as detected by the above statistical criterion) and that they intervene in a sufficiently large number of interfaces. The number of interfaces in which a given element X takes part is called the *coverage* of the element and is noted $cov(X)$. A single threshold on the minimal coverage of elements of interest will select the elements that play a role in at least that many interfaces or fraction of the interfaces. For instance, in our study, we chose a 5% threshold for the minimal coverage of elements to be considered for further analysis.

[4] The term *semi-interface* (possibly associated with a subscript) denotes the half belonging to one protein in an interface.

[5] Strictly speaking, the probability of measuring n_X^{obs} outside the range $[\mathbb{E}[n_X] \pm 1.96\sqrt{Var[n_X]}]$ is less than 5%. For symmetry reasons, $p\big(n_X^{obs} \geq \mathbb{E}[n_X] + 1.96\sqrt{Var[n_X]}\big) < 2.5\%$. This is also known as the p-value.

[6] For instance, if one keeps all items with $n_X^{obs} \geq \mathbb{E}[n_X] + \sqrt{Var[n_X]}$, then there is a 16% chance that this happened under the null hypothesis.

2.4 Measuring the spread of the elements and its atypical character

It is also informative to know if a given element (an item or a combination of items) tends to occur in a widespread fashion among the interfaces or, on the contrary, in a concentrated way. In the former case, this might indicate a necessary ingredient in at least one type of bonds between proteins. In the latter case, this could be interpreted as the sign of a kind of autocatalytic reaction that favors the co-occurrence of a same element inside interfaces. Either way, one must be able to measure to which degree an element is more widespread or more concentrated than normal. We therefore propose to *compare the measured coverage of elements with their expected coverage*.

The coverage of an element is easily computed from the database of known instances. The computation of its expected coverage, on the other hand, requires some caution. Suppose that a given element has been observed to occur n times. The idea is to calculate the number of different interfaces among I (e.g. 459) that can receive at least one element when n elements of the same type are drawn independently within N elements.

Suppose that the average number of elements in each interface is $K = N/I$, and let k be the number of a given element in a given interface. Then k is the size of the intersection between a set of n elements drawn independently from N and a set of K elements also drawn independently from N. The hypergeometrical equation gives:

$$p(k) = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}$$

In particular, the probability of having a void interface (w.r.t. the element of interest) is:

$$p(0) = \frac{\binom{n}{0}\binom{N-n}{K-0}}{\binom{N}{K}} = \frac{\binom{N-n}{K}}{\binom{N}{K}}$$

And the expected number of *non void* interfaces is:

$$I \cdot \overline{p(0)} = I \cdot \left(1 - \frac{\binom{N-n}{K}}{\binom{N}{K}}\right)$$

3 Results on the protein-protein interfaces

Item selection

459 protein-protein complexes were taken from the PDB database [7,2], the existing total number at the time of the first experiments, and have been described as explained above using 120 different items. For each of these items, the total number of occurrences and the coverage have been measured, while the expected number of occurrences (with standard deviation) and the expected coverage have been computed. The `aCID` method then automatically extracted the items satisfying the three criteria:

- *Overrepresentation*. $n_X^{obs} \geq \mathbb{E}[n_X] + C\sqrt{Var[n_X]}$, with $C = 2$ when selecting items, and $C = 1$ or $C = 2$ when selecting patterns.
- *Minimum coverage*. A threshold of 5% or 7% was used for the selection of patterns. None was used when selecting items.
- *Difference with expected coverage*: $cov(X) > \mathbb{E}[cov(X)]$.

+-	SSP	SHP	SPP	SP-	S+-	HHH	HHP	HH+	HP+	H+-	PP+	P+-	++-

Table 1. Selected items

It is noteworthy (see table 1) that items known as poor candidates such as: $--$, $++$, $---$, $+++$ have been rejected. On the other hand, items corresponding to mildly hydrophobic or strongly hydrophobic elements have been retained, such as HHH, HHP, $HH+$, as well as electrically charged elements such as $+-$, $S+-$, $H+-$, $P+-$, $++-$. All of these items are indeed expected to play a role in protein-protein interfaces since they tend to favor stable conformations.

Pattern selection

The same analysis was carried over for the combinations of items, including *doublets*, *triplets*, and *quadruplets* (no *quintuplet* were found to satisfy the selection criteria). In order to test the robustness of the results, selection was carried out using $C = 2$ and $C = 1$ for the overrepresentation criterion and a minimal coverage threshold of 5%.

Doublets	$\boldsymbol{SSP/SSP}$, $\boldsymbol{SSP/SPP}$, $\boldsymbol{SP-/SP-}$, $\boldsymbol{S+-/S+-}$, $\boldsymbol{S+-/++-}$, $\boldsymbol{HHH/HHH}$, $\boldsymbol{HHH/HHP}$, $\boldsymbol{HHH/HH+}$, $\boldsymbol{HHH/H+-}$, $\boldsymbol{HHP/HHP}$, $\boldsymbol{HHP/HP+}$, $\boldsymbol{HH+/HH+}$, $\boldsymbol{HP+/HP+}$, $\boldsymbol{H+-/H+-}$, $\boldsymbol{H+-/++-}$, $+-/+-$, $+-/++-$, SHP/SHP, $S+-/P+-$, $HHP/HH+$, $HH+/HP+$, $HH+/P+-$, $HH+/++-$
Triplets	$\{\boldsymbol{S+-/S+-/S+-}$, $\boldsymbol{++-/H+-/S+-}$, $\boldsymbol{HH+/HHH/HHH}$, $\boldsymbol{HH+/HH+/HHH}$, $\boldsymbol{H+-/H+-/H+-}$ $\{+-/+-/+-$, $+-/HH+/HH+$, $+-/H+-/H+-$, $SHP/SPP/SSP$, $SHP/SHP/SHP$, $H+-/H+-/S+-$, $HH+/HHH/HHP$, $H+-/HHH/HHP$, $H+-/HH+/HHH$, $H+-/H+-/HH+$, $H+-/H+-/HP+$, $H+-/HH+/HH+\}$
Quadruplets	$\{\boldsymbol{H+-/HH+/HHH/HHH}$, $+-/HH+/HHH/HHH$, $SHP/SPP/SSP/SSP$, $SHP/SHP/SHP/SHP$, $++-/H+-/S+-/S+-$, $H+-/HH+/HH+/HHH$, $HHP/SHP/SHP/SHP$, $HH+/HHH/HHP/HHP$, $H+-/HHH/HHP/HHP\}$

Table 2. Items that are over-represented (C=2 (bold), and C=1) and cover at least 5% of the interfaces. Results for C=1 are a superset of the results for C=2.

Regarding the *doublets* and the *triplets*, one can notice a slight overrepresentation of the groups with amino acids belonging to the hydrophobic group H. In general, however, the items are paired according to global properties. Two groups of patterns emerge. One with a high proportion of hydrophobic amino acids H, the other with opposite charges + and -. Only in one instance these properties are found together: $HH+/++-$.

As for the *quadruplets*, it is noticeable that hydrophobic amino acids are predominant. The electric charges + and - equilibrate each other, and there is a positive charge + left. The groups with hydrophobic amino acids takes over.

4 Discussion and future work

Protein docking introduces very challenging problems. In this work, we used a low-resolution geometrical description of protein-protein interfaces and a data mining approach combined with a null hypothesis criterion. This discovery method can be applied in every contexts where the representation of the instances involves counts of patterns taken in a dictionary that is not too large wrt. the number of instances. Furthermore, it naturally adapts to the discovery of disjunctive concepts i.e. different causal processes. Finally, there is no need for constructing artificial decoys.

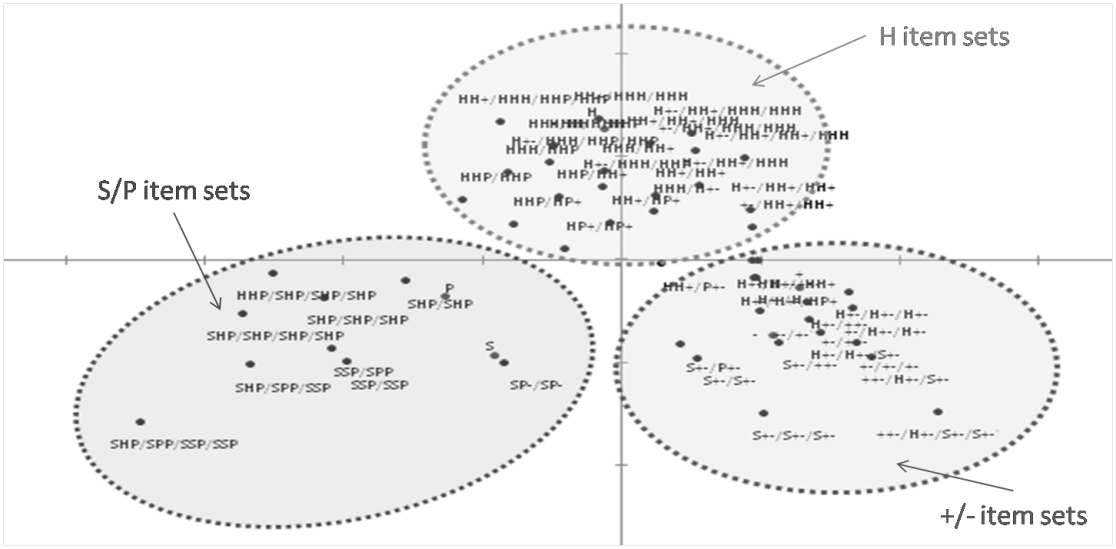

Figure 2. Results of the PCA analysis of the extracted item sets.

Applied to the data set of 459 protein-protein complexes taken from the Dockground database, the `aCID` method selected items and the combinations thereof that point out to the importance of the hydrophobic amino acids and the association of amino acids of opposite charges. The findings are aligned with what is known about protein-protein complexes. Moreover, the results are robust against variations in the grouping of the amino acids into five groups (S, P, H, + and -) and changes in the threshold for selecting significant patterns. The value of the alpha parameter for the alpha-shapes should, however, play a much more important role. This remains to be systematically studied.

References

[1] R.P. Bahadur, P. Chakrabarti, F. Rodier and J. Janin, A Dissection of Specific and Non-specific Protein-Protein Interfaces, *J. Mol. Bio.*, 336, 2004.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, N. Shindyalov and P. E. Bourne, The protein data bank, *Nucleic Acids Research (NAR)*, 28:235-242, 2000.

[3] J. Bernauer, J. Azé, J. Janin and A. Poupon, A new protein-protein docking scoring function based on interface residue properties, *Bioinformatics*, 23:555-562, 2005.

[4] M. Betts and R. Russell, Amino acid properties and consequences of substitutions, *Bioinformatics for Geneticists*, M.R. Barnes, John Wiley and Sons, pp. 291-315, 2003.

[5] F. Cazals, J. Giesen, M. Pauly and A. Zomorodian, Conformal Alpha Shapes, *In proceedinfs of Eurographics Symposium on Point-Based Graphics*, 2005.

[6] P. Chakrabarti and J. Janin, Dissecting protein-protein recognition sites, *Proteins: Structure, Function, and Bioinformatics*, 47:334-343, 2002.

[7] D. Douguet, H.-C. Chen, A. Tovchigrechko and I. A. Vakser, Dockground resource for studying protein-protein interfaces, *Bioinformatics*, Oxford University Press, Oxford, UK, 22:2612-2618, 2006.

[8] H. Edelsbrunner, Weighted alpha shapes, Technical report, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 1992.

[9] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym and G. Schreiber, From The Cover: The modular architecture of protein-protein binding interfaces, *PNAS*, 102:57-62, 2005.

[10] G. R. Smith and M. J. Sternberg, Prediction of protein-protein interactions by docking methods, *Curr Opin Struct Biol.*, Feb, vol. 12, num. 1, pp. 28-35, 2002.