Identifying interface elements implied in protein-protein interactions using statistical tests and Frequent Item Sets

Christine Martin
LIMSI CNRS,Bt 508,
Université d'Orsay Paris Sud
BP 133, 91403 Orsay cedex (France)
martinc@limsi.fr

Antoine Cornuéjols
UMR AgroParisTech/INRA 518
AgroParisTech
16 rue Claude Bernard, 75231 Paris cedex 05 (France)
antoine.cornuejols@agroparistech.fr

Abstract

Understanding what are the characteristics of protein-protein interfaces is at the core of numerous applications. This paper introduces a method in which the proteins are described with surfacic geometrical elements. Starting from a database of known interfaces, the method produces the elements and combinations thereof that are characteristic of the interfaces. This is done thanks to a frequent item set technique and the use of statistical tests to ensure a marked difference with a null hypothesis. This approach allows one to easily interpret the results, as compared to techniques that operate as "black-boxes". Furthermore, it is naturally adapted to discover disjunctive concepts, i.e. different underlying processes. The results obtained on a set of 459 protein-protein interfaces from the PDB database confirm that the findings are consistent with current knowledge about protein-protein interfaces.

1 Introduction

Understanding how proteins function, and, particularly protein-protein interactions, is of paramount importance for biology and medicine and is a coveted prize for many research efforts [13]. Being able to predict interactions would permit searching a database of proteins and retrieve all proteins that can interact with a given molecular structure, e.g. another protein [4, 11]. This entails two sub-questions: *how to predict that two proteins may interact* (i.e. bind) and, if so, *what are the most likely docking sites* (i.e. the ones associated with the minimal energy of the compound structure)?

Several directions have been explored to solve these problems. The first one is to *start from fundamental premisses*, meaning here 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 try to *learn by automatic means to discriminate native-like protein-protein complexes from decoys*. Many questions then arise. One concerns the choice of the parameters used to represent proteins and protein-protein complexes that are input to the learning algorithm. Another question is the determination of negative instances of complexes, or decoys. As, usually, the databases only contain positive instances of protein-protein complexes, negative instances have to be generated. Most often, this consists in generating other, random, conformations and orientations of the proteins, assuming that these correspond to bad or impossible pairings. However, the quality of these negative instances may be disputed and can profoundly affect the performance of the learning methods.

This is why another approach is proposed in this research work. 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 techniques of data mining in order to detect characteristic regularities in known protein-protein interfaces. This approach presents three advantages over the use of standard supervised machine learning techniques. First, there is no need for building questionable negative instances of protein-protein complexes. Second, the results can be analyzed and may lead to a better understanding of the formation of complexes. And, third, it is easy to convert the regularities discovered into a prediction tool that scores the potential pairings of proteins.

The paper starts with the representation used for the description of the protein structure. Section 3 describes in a generic way the new proposed method to analyze whether the interfaces of protein-protein complexes have special properties, and, if yes, which ones. The results obtained using the Protein Data Bank [3] and data extracted from the Dockground database [9] are described in section 4. Finally, section 5 discusses the results obtained in light of the overall protein docking problem and opens directions for extensions of this work.

2 Representing the geometry of proteins

In order to understand how two proteins can form a complex, we need both to know what defines their surfaces, and more precisisely their interface, and which atoms or amino acids participate in this interface [2, 12]. In this work, we rely on a weigthed alpha complex model which, under some constraints, can be considered as a dual of the Voronoï model [8].

Let us consider the set of balls $B = B_i(p_i, w_i)$, where p_i represents the center of ball i and w_i the square of its radius. Then, the *weighted alpha shape* is the domain of these balls and the *weigthed alpha complex* is the corresponding skeleton, with alpha controlling the desired level of precision [7].

It contains geometrical patterns (strictly speaking *simplexes*, later to be called *items*) that are: the *points* of B, the *edges* defined by the centers of pairs of balls that have a non empty intersection, the *triangle* and the *tetrahedron* defined analogously. The figure 1 shows an example of this representation in two dimensions, with 14 spheres or balls, with a null alpha (a) and with a non null alpha (b). Alpha controls the level of detail by varying the radius of balls ($\sqrt{w_i + \alpha}$) and, hence, the nature of the geometrical representation. For more details, see [10].

Equating amino acids with balls (with a radius depending on their physico-chemical properties), the interface of a protein-protein complex is defined as the set of *edges*, *triangles* and *tetrahedra* of which at least one point belong to each of the paired proteins.

Additional information about the nature of the amino acids involved is needed. However, because there are twenty different amino acids, the number of possible elementary geometrical items (edges, triangles and tetrahedra) is too large ($\approx$ 10600) compared to the number of available examples (here 459 selected complexes). Hence, we have decided to group the amino acids with respect to their physico-chemical properties [6]. Considering five groups defined according to broad biochemical properties: hydrophobic, polar, positively charged, negatively charged and small (resp. noted H, P, $+$, $-$ and S), thus brings down the repertoire to 120 distinct descriptive items.

Using this representation, one typical interface between proteins involves between 15 to 50 items, some of them possibly repeated.

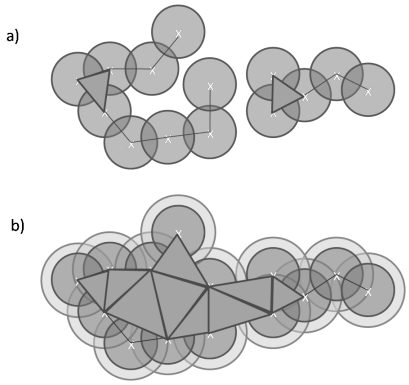

Figure 1. (a) the alpha shape in 2D with alpha=0, and (b) with alpha$\neq$ 0.

3 Analysis by frequent item sets and statistical hypothesis testing

Thanks to the method described in the previous section, 459 protein-protein complexes from the Protein Data Bank [3] have been represented as well as their interfaces[1]. Each one of these interfaces is characterized by a collection of descriptors taken out from 120 elementary geometrical items. For instance, a given interface could be described as: $\langle (S_1 \leftrightsquigarrow +_2), (S_1 H_1 -_1 \leftrightsquigarrow +_2) \rangle$, where the subscripts are used to indicate which protein (either protein 1 or protein 2) is providing the amino-acid. In this example, the interface involves two elementary geometrical items: an *edge* made of a small amino-acid S coming from protein 1 and a positively charged amino-acid $+$ coming from protein 2, together with a *tetrahedron* of which the triplet $(SH-)$ is provided by protein 1 and the amino-acid $+$ comes from protein 2. On average, the interfaces are composed of approximately 22 geometrical items, which amounts to approximately 10,000 items for the whole set of 459 interfaces.

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

Given the relatively small numbers of known interface components ($\approx$ $10,000$) as compared to the number of descriptors (120), the regularities that can be reasonably looked for are, per force, of limited complexity. One first question relates to the composition of the interfaces in terms of items. Is this composition special in some ways?

[1] This was done using the CGAL librairy [1].

3.1 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 this question 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. The question is akin to ask what *should* be the number of each item given the surface composition of the proteins making the compound. In the following, the term *semi-interface* (possibly associated with a subscript) denotes the half belonging to a protein in an interface.

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[2] is:

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

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:

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

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

One important caveat in our case is 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] I.e. independent and identical trials.

3.2 Analysis of the combinations of items

It can be expected that what determines the strong coupling between the complex structures of proteins are combinations of elementary geometrical items. It is therefore interesting to 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 $\mathtt{A}$ of m items $\underbrace{\mathtt{A}_i, \mathtt{A}_j, \ldots, \mathtt{A}_k}_{m}$ is:

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

and the variance:

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

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

3.3 Combing the items and combinations

Underrepresented items, while of possible predictive value (their absence in an candidate interface could be considered as a telltale sign of a possible interface), 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% [3]. This rather conservative threshold contributes to yield few false positive elements. [4].

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.

[3] 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.

[4] 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.

3.4 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.

To do this, we 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 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:

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

Several cases can occur that deserve to be considered:

1. The *number of observed elements is **greater** than the number of expected elements* AND the *observed coverage is **larger** than the expected one*. This means that the element is over-represented and is overspread in the known instances. It might therefore be thought of as playing a key role in the binding of proteins.

2. The *number of observed elements is **greater** than the number of expected elements* AND the *observed coverage is **less** than the expected one*. This means that the element is over-represented and concentrated on some interfaces. This may suggest a kind of autocatalytic process involving the element.

3. The *number of observed elements is **less** than the number of expected elements* AND the *observed coverage is **larger** than the expected one*. The element is under-represented and is overspread in the known instances. It might be accidentally associated with the interfaces.

4. The *number of observed elements is **less** than the number of expected elements* AND the *observed coverage is **less** than the expected one*. The element is under-represented and concentrated on some interfaces.

4 Results on the protein-protein interfaces

4.1 Item selection

459 protein-protein complexes have been extracted from the PDB database, and have been described using weighted alpha shapes (with alpha=0), corresponding to approximately 10,600 items (edges, triangles and tetrahedra). In this description, each interface has been found to involve between 15 to 50 items. There are 120 different items when the amino acids are clustered into five groups. For each one 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 (see the bar chart in figure 2).

We then automatically extracted the items that satisfied a combination of the three criteria:

- *Overrepresentation*.
  When $n_X^{obs} \geq \mathbb{E}[n_X] + C\sqrt{Var[n_X]}$, 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*.
  When $cov(X) > \mathbb{E}[cov(X)]$.

The list of selected items is then: $\{+-, SSP, SHP, SPP, SP-, S+-, HHH, HHP, HH+, HP+, H+-, PP+, P+-, ++-\}$.

It is noteworthy 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.

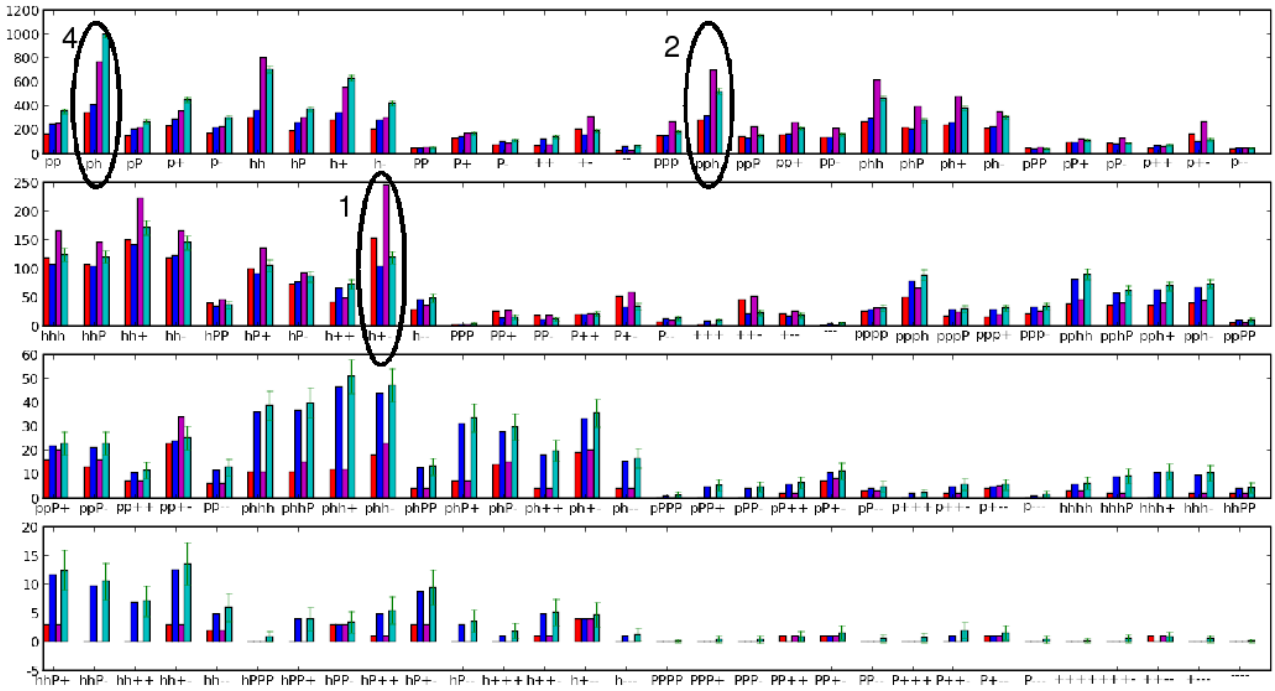

Figure 2. Bar charts for the items. For each item, the columns are as follows. First: *Observed coverage*. **Second:** *Expected coverage*. **Third:** *Observed number of occurrences*. **Fourth:** *Expected number of occurrences* **with a standard deviation bar. The ellipses point to instances corresponding to the classes of interesting cases described in section 3.4.**

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 1. Items that are over-represented (C=2, results in 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.

4.2 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, another selection was also carried out using $C = 1$ for the overrepresentation criterion and a minimal coverage threshold of 5%.

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.

Calculations were also carried out imposing a minimal coverage threshold of 7%, but with the less stringent condition $C = 1$. The results show a large agreement with the ones obtained for $C = 2$, except for the triplets. A finer analysis is under way to look whether this discrepancy is profound or not.

Doublets:
$\{+-/+-, SSP/SSP, SHP/SHP, S+-/S+-,$
$HHH/HHH, HHH/HHP, HHH/HH+, HHH/H+-,$
$HHP/HH+, HHP/HP+, HH+/HH+, HH+/HP+,$
$H+-/H+-, H+-/++-\}$

Triplets:
$\{+-/H+-/H+-, SHP/SHP/SHP, H+-/H+-/S+-,$
$H+-/HH+/HHH, H+-/H+-/HH+\}$

Quadruplets: $\{H+-/HH+/HHH/HHH\}$

5. Discussion and future work

Protein docking introduces very challenging problems. In this work, we relied on a low-resolution geometrical description of protein-protein interfaces. By contrast with the usual scoring functions that rely on aggregations of multiple factors, the method we propose searches for (geometrical) patterns that emerge as strongly correlated with protein-protein interfaces. One important advantage of this approach is that it naturally adapts to the discovery of disjunctive concepts. While ordinary methods may be severely hampered in their performance by the fact that the phenomenon at hand might in fact result from several different processes, our method is geared to bring to light such compound models (for instance, antibody, enzymes, cytokine, and other interaction classes). Another advantage is that there is no need for constructing artificial decoys (e.g. [5]). On the other hand, null hypotheses have to be devised in order to test the significance of the number of occurrences of each pattern and of their coverage.

This discovery method can be applied in every context where the representation of the instances involves counts of patterns taken in a dictionary of patterns that is not too large as compared to the number of instances.

It is not difficult to turn the discovered patterns into a predictive tool. It suffices to retain the combinations of patterns that are the most statistically significantly associated with the phenomenon and that together cover most or all of the positive known instances. Each combination can then be used as a predictor of potential interfaces in protein-protein complexes.

We applied our method to the data set of 459 protein-protein complexes taken from the Dockground database. The items and the combinations thereof that were selected 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.

We tested the robustness of the results against variations in the grouping of the amino acids into five groups (S, P, H, + and -) and in the threshold for selecting significant patterns. The results stayed qualitatively the same, with more patterns selected when the selection threshold was made less severe. The value of the alpha parameter for the alpha-shapes should, however, play a much more important role in the kind of patterns discovered by the method. This remains to be systematically studied.

Acknowledgements. We thank warmly Ludovic Autin, Julie Bernauer and Frederic Cazals for their insightful comments on an earlier version of this text.

References

[1] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[2] R. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Bio.*, 336, 2004.

[3] 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(1):235–242, 2000.

[4] J. Bernauer. *Utilisation de la tessellation de Vorono pour la modlisation des complexes protine-protine*. Ph.d. thesis, Universit Paris Sud XI, 2003.

[5] J. Bernauer, J. Az, J. Janin, and A. Poupon. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5):555–562, 2005.

[6] M. Betts and R. Russell. Amino acid properties and consequences of substitutions. In M. Barnes, editor, *Bioinformatics for Geneticists*, pages 291–315. John Wiley and Sons, 2003.

[7] F. Cazals, J. Giesen, M. Pauly, and A. Zomorodian. Conformal alpha shapes. In *Eurographics Symposium on Point-Based Graphics*, 2005.

[8] F. Cazals, F. Proust, R. Bahadur, and J. Janin. Revisiting the voronoi description of protein-protein interfaces. *Protein Science*, 15(9):2082–2092, 2006.

[9] D. Douguet, H.-C. Chen, A. Tovchigrechko, and I. A. Vakser. Dockground resource for studying protein–protein interfaces. *Bioinformatics*, 22(21):2612–2618, 2006.

[10] H. Edelsbrunner. Weighted alpha shapes. Technical report, Champaign, IL, USA, 1992.

[11] M. F. Lensink, R. Mndez, and S. J. Wodak. Docking and scoring protein complexes: Capri 3rd edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4):704–718, 2007.

[12] 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(1):57–62, 2005.

[13] G. Smith and M. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol.*, 12(1):28–35, Feb 2002.