Frequent itemset mining on protein-protein interfaces

C. Martin[1,2] and A. Cornuéjols[2]

[1] LIMSI CNRS,Bt 508,Université d'Orsay Paris Sud BP 133,91403 Orsay cedex
martinc@limsi.fr
[2] UMR AgroParisTech/INRA 518, AgroParisTech, 16 rue Claude Bernard, 75231 Paris cedex 05
antoine.cornuejols@agroparistech.fr

Keywords: protein-protein docking, frequent itemset mining.

1 Introduction

Many important biological processes involve protein-protein interactions. These correspond to contacts, also called dockings, on an interface area (from 900 A^2 to 2000 A^2). Numerous research efforts have been aimed at predicting these interactions, using either automatic or immersive methods. Neither of these two families of techniques is completely satisfactory however [1]. On one side, *automatic methods* are very costly in terms of computation because the large number of degrees of freedom in the representation of the phenomenon leads to a gigantic search space, even when using discretized attributes. Moreover, the current knowledge of the forces involved in these interactions is still scarce and uncertain and, as a result, evaluation functions used to guide the search process are ill-informed, leading to the exploration of numerous dead ends and yielding many false positive solutions. On the other side, *immersive methods*, that seek to take advantage of the expert's "intuition" about protein docking, specially with regards to geometric configurations, are limited because some feedback devices, like haptic ones, require real-time computations of forces that are difficult to obtain, while, at the same time, identifying the most useful display of additional information needed by the expert is still a matter of research in virtual reality environments.

This is why a new hybrid approach (combining immersive and automatic context) named HOSMoS (Human Oriented Selection of Molecular Specimen) has been proposed [1]. The idea is to rely as much as possible on the expert's knowledge in order to search and evaluate the possible docking situations while easing his/her task by providing predictive features that can be computed in real-time about tentative dockings. The problem is therefore to find relevant and easy to compute "abstractions" that help in the evaluation of the possible solutions. This paper presents a method for the identification of such features in the complex context of protein-protein interactions.

2 The identification of relevant features

The central problem we face in the study of protein-protein interactions is the identification of predictive features. One method, in this context, is to use supervised learning in order to select the features that allow learning algorithms to discriminate between positive and negative instances. However, in the case of protein-protein interactions, only positive instances are known, and in limited amount at that. Putative negative instances may possibly be generated, but their information content would be of very diverse and uncontrolled value. Another line of attack is then to look for conjunctions of descriptors (called *itemsets*) of which the rate of appearance in the known positive instances significantly differs from the rate one would expect given no information about the class. For instance, one could find that the conjunction of descriptors a_1 & a_6 & a_{23} is present in 20% of the positive instances, while one would expect it *a priori* in a proportion of only 1%.

This approach requires solving two problems. First, a dictionary of descriptors must be identified, which is both informative about the phenomenon at hand, but also limited enough so as to enable one to get statistically significant counts of conjunctions of these in the positive examples. Second, a technique must be found in order to compute *a priori* expected rates of itemsets, something also known as a "null hypothesis". Before describing our adaptation of the frequent item set method to our problem, we first present the data representation scheme we use.

3 Data selection and representation

Data about macromolecular structures like protein-protein complexes are getting increasingly available. However, quality and redundancy issues require one to be selective. In our case, 460 protein-protein complexes from the Dockground [3] database were retained. From biological considerations, it is suggested that the contacts between the amino acids involved in the interfaces are a determining factor. We have therefore resorted to a model [4] based on three geometrical objects, namely, *edges*, *triangles* and *tetrahedra* (Figure 1) coding the geometry of a set of spheres, each one representing an amino acid. Edges represent the contact of two amino acids, triangles and tetrahedra the contact of three (resp. four) amino acids. Using properties of triangulation in a three dimensionnal space, these three geometrical items are sufficient to describe a set of spheres in a unique way.

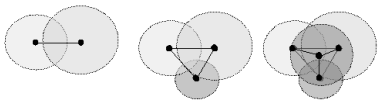

Fig. 1. Elementary objects of the used protein representation

Because there are twenty different amino acids in nature, the number of possible items (approx. 10'600) is too large compared to the number of available examples. Hence, we have decided to group the amino acids with respect to their physico-chemical properties. Taking five groups thus leads to a repertoire of 120 distinct descriptive items. Using this representation, one typical interface between proteins involves between 15 to 50 items, some of them possibly repeated.

4 Experiments and perspectives

It is then easy to compute the number of times each item appears in all positive instances. It remains to be decided which items (if any) are predictive of a potential docking. One difficulty is that classes of items (edges, triangles, tetrahedra) appear naturally with different frequencies (e.g. tetrahedra are less likely to appear than edges), requiring that specific decision thresholds be determined. This is a rather well-known problem in frequent item sets mining methods [5]. In our case, we decided to compute a base probability for each item, corresponding to the probability of the item to be part of an arbitrary interface (positive or not). The resulting analysis, still in progress, has, on one hand, filtered out items that were known to be good candidates for complementary roles in docking, and, on the other hand, pointed out unexpected patterns that could be useful indices of potential dockings. For instance, some tetrahedron comprising the same four groups of amino acids is more frequently found than triangles of the same groups!

It therefore appears that our approach for finding relevant and easy to compute description features can both help understanding the domain at hand and provide useful components for the definition of the evaluation functions required to guide the search in an otherwise gigantic and complex state space.

References

1. N. Ferey, J. Nelson, G. Bouyer, C. Martin, P. Bourdot, and J.-M. Burkhardt, User needs analysis to design a 3d multimodal protein-docking interface, In IEEE 3DUI 2008, 8-9th March, Reno, Nevada, USA, 2008.
2. P. Chakrabarti and J. Janin, Dissecting protein-protein recognition sites Proteins: Structure, Function, and Bioinformatics, 47 (3), 334-343, 2002.
3. D. Douguet, H.-C. Chen, A. Tovchigrechko and I. A. Vakser, DOCKGROUND resource for studying protein-protein interfaces, Bioinformatics, 22(21), 2612-2618, 2006.
4. F. Cazals, J. Giesen, M. Pauly and A. Zomorodian, Conformal Alpha Shapes, Eurographics Symposium on Point-Based Graphics, 2005.
5. B. Liu, W. Hsu and Y. Ma, Mining Association Rules with Multiple Minimum Supports, ACM SIGKDD, August 15-18, San Diego, CA, USA, 1999.