A Template-Matching Approach for Protein Surface Clustering

Abstract

Surface-based techniques for protein comparison and classification typically require a compact surface representation, capable of effectively condensing its description. In this paper we propose an original template-matching algorithm for multi-feature surface clustering in the biochemical context. The effectiveness of our clustering algorithm in capturing surface similarities is then discussed within a larger framework for protein classification based on surface comparison, with the support of tests performed on a dataset including 25 proteins.

1. Introduction

Understanding which characteristics of proteins have most impact on their functional role is one of the main challenges of the post-genomic era. In this direction, since protein functions occur predominantly on or near the protein surface, comparison of protein surfaces has been gaining more and more significance in the last decade in an attempt to discover functional relationships not revealed by techniques for structural comparison. In fact protein comparison, that consists in determining similar surface portions on different molecules, may sometimes point out different structures or sequences coming together in a unique active site having a common function [3][4].

The choice of the surface description is doubtlessly critical, since a satisfactory trade-off should be reached between the representation grain and the performance of the comparison algorithm. Mainly three approaches have been devised to describe the molecular surface, namely *mesh-based*, *atom-based*, and *patch-based*, briefly discussed in the following.

Most mesh-based methods use the so-called Connolly algorithm [2] to obtain a mesh-based representation of the solvent accessible surface. On the other hand, atom-based methods define the molecular surface as a significant subset of the solvent exposed atoms. For example, in [10] the author defines the α-surface as the set of atoms touched by a probe of given radius and adopts this description to discover surface patterns within a database of proteins aimed at classifying them.

In patch-based methods, the grain of the surface is increased by considering the set of exposed amino acids or a compact representation obtained by segmenting the surface into homogeneous regions. A graph-based method for protein comparison and classification using a patch representation of the surface has been described in [7], where comparison relies on matching surface graphs whose nodes are concave, convex, and toroidal patches extracted from the Connolly surface.

In the context of patch-based methods, in this work we address the problem of defining a compact surface representation for a protein, capable of effectively condensing the description of its local properties. To this end, after briefly discussing the specific requirements raised by protein comparison, we propose a multi-feature clustering technique, based on a template-matching algorithm, that starting from the punctual 3D description of a protein surface generates a set of homogeneous patches. In evaluating homogeneity, differently from other approaches that only take into account geometrical properties of the protein surface (e.g., concavity and convexity) [9][7], we also consider its electrostatic potential, that plays a very relevant role during protein interactions. We believe that the compact representation obtained this way for protein surfaces is an excellent starting point for developing a robust technique for surface-based protein comparison. To prove this claim, we test our clustering technique within the surface-based classification framework proposed in [1], on a dataset including 25 proteins.

The rest of the paper is structured as follows. In Section 2 we briefly discuss how clustering is framed within our approach to surface-based protein classification; in Section 3 we describe our template-matching approach to surface clustering; in Section 4 we present the classification test we carried out; finally, in Section 5 we discuss the validity of the approach.

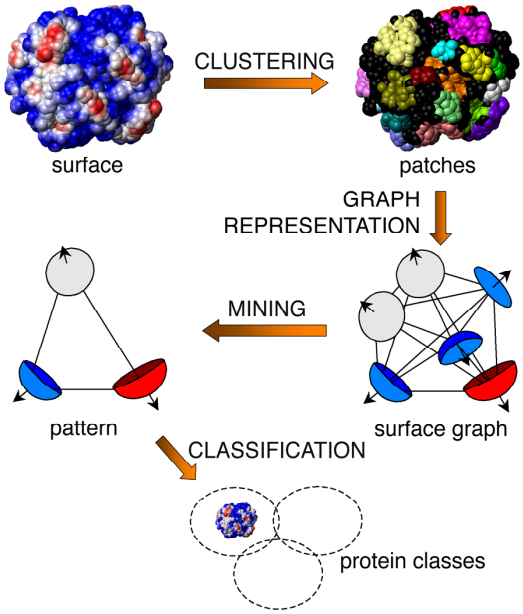

Figure 1. Surface-based classification

2. A Framework for Clustering

In this section we summarize the framework proposed in [1] for classifying proteins based on their surface properties. The classification results will be used in Section 4 to indirectly evaluate clustering, since the soundness of the classification obtained for a dataset of known proteins depends on how effectively clustering represents their surfaces.

As sketched in Figure 1, the classification approach consists of four steps:

1. *Clustering*: determine, on each protein, a set of homogeneous and connected surface regions (*patches*). Details are given in Section 3.
2. *Graph representation*: synthetically represent each protein by a spatial graph of patches (*surface graph*). In fact the protein properties do not only depend on the set of patches characterizing the surface, but also on their relative positioning and orientation.
3. *Mining*: find frequent patterns of patches in the protein dataset through mining techniques. A *pattern* on a protein is a subgraph of its surface graph, thus it models a set of patches and their relative spatial placement. In order to mine frequent patterns, we developed a level-wise algorithm that iteratively determines frequent patterns made up of an increasing number of patches.
4. *Classification*: classify proteins into a dendrogram, by grouping together proteins that share common frequent patterns. We adopt a hierarchical technique that, starting from clusters composed by a single protein, progressively merges the two most similar ones according to the complete-link approach. Similarity is based on the number of shared patterns; the larger the pattern, the higher the contribution to the score function.

3. Surface Clustering

Clustering is applied starting from a representation of the surface of a protein p in form of a connected, non-directed graph $G_p = (V_p, E_p)$ where nodes $v \in V_p$ represent surface atoms of p and E_p includes the edges (v_i, v_j) such that atoms v_i and v_j are adjacent on the surface of p. Each node v is labeled with its coordinates in the 3D space and with the local values of surface curvature, $cur(v)$, and electrostatic potential, $elp(v)$.

In turn, this coarse-grained representation derives from a fine-grained one given in terms of triangular meshes. The local curvature in v is estimated on the discrete surface formed by its adjacent nodes, according to the method for the surface variation calculation proposed in [8]. The electrostatic potential is computed by the MOLMOL program [5].

3.1 Requirements

The main requirements for protein surface clustering, as emerging from the application domain, can be summarized as follows:

1. Each patch should be connected, not overlapping with the others, and should cover at least 10 surface atoms in order to enable significant inter-protein interactions.
2. Each patch should present homogeneous values for both curvature and potential.
3. Patches should have a regular shape in order to better characterize specific regions of the protein.
4. The union of the patches should cover a significant percentage of the protein surface.

In the light of these requirements, we propose a clustering technique that determines an optimal (in terms of both homogeneity and covering) assignment of nodes (i.e., atoms) to patches whose shape resembles that of a given regular template. In particular, we adopted a circular template defined as a patch c with center on node v and integer radius r; all the nodes whose distance from v is less than r edges belong to patch c.

3.2 Features

The application domain requires that patch homogeneity is evaluated on the basis of a proper feature categorization: three categories are relevant for the curvature (convex, planar, concave), three categories

for the electrostatic potential (negative, neutral, and positive). Since there is no evidence of a sharp separation between these categories, we smooth the discretization by introducing parametric grey areas as depicted in Figure 2.

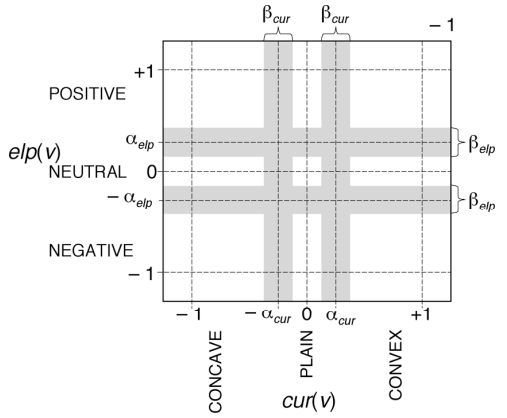

Figure 2. Feature discretization

The discretization is ruled by two parameters, α_f and β_f, that respectively determine the center and the width of the grey area between two categories for feature *f*. Based on this discretization, the category(ies) for node *v* are defined as follows:

$$\hat{cur}(v) = \begin{cases} concave, & \text{if } cur(v) < -\alpha_{cur} + \frac{\beta_{cur}}{2} \\ plain, & \text{if } -\alpha_{cur} - \frac{\beta_{cur}}{2} < cur(v) < \alpha_{cur} + \frac{\beta_{cur}}{2} \\ convex, & \text{if } cur(v) > \alpha_{cur} - \frac{\beta_{cur}}{2} \end{cases}$$

$$\hat{elp}(v) = \begin{cases} negative, & \text{if } elp(v) < -\alpha_{elp} + \frac{\beta_{elp}}{2} \\ neutral, & \text{if } -\alpha_{elp} - \frac{\beta_{elp}}{2} < elp(v) < \alpha_{elp} + \frac{\beta_{elp}}{2} \\ positive, & \text{if } cur(v) > \alpha_{elp} - \frac{\beta_{elp}}{2} \end{cases}$$

The same discretization is used to define the category of patch *c* according to feature *f*, as the category of the average value of *f* for the nodes in *c*. The nodes (and the patches) for which a single category is returned for each of the two features are said to be *white*.

A homogeneity index can be defined for the patches, in order to evaluate the degree of homogeneity of its nodes with respect to each feature. A node *v* is said to be *compatible* with a white patch *c* on feature *f* iff $\hat{f}(c) \subseteq \hat{f}(v)$, i.e., if either their category is the same, or *v* falls into a grey area adjacent to the category of *c*. The homogeneity for feature *f* of patch *c* including *m* nodes is then defined as:

$$hom(c, f) = \frac{\#f}{m}$$

where $\#f$ is the number of nodes in *c* that are compatible with *c* on feature *f*.

3.3 Algorithm

The template-matching algorithm consists of two steps:

(1) Definition of Candidate Patches. The set C_{cand} of candidate patches consists of all the white patches *c* with radius *r* lower than r_{max} for which $hom(c, f)$ is higher than a threshold σ for each *f*.
(2) Optimization. An optimal set of non-overlapping patches is selected from C_{cand} by exactly solving a pure binary integer programming problem; optimality is defined by encouraging large patches.

Let $t = \#C_{cand}$ be the number of candidate patches and m_i be the number of nodes in candidate patch c_i; the integer formulation is as follows:

$$\begin{aligned} &Maximize: && \sum_{i=1}^{t} m_i^2 x_i \\ &Subject\ to: && x_i \in \{0,1\} \qquad \text{for } i = 1,..,t \\ & && x_i + x_j \le 1 \quad \forall c_i, c_j ; c_i \cap c_j \neq \varnothing \end{aligned}$$

Each binary variable x_i has value 1 if patch c_i is in the solution, 0 otherwise. The constraint is aimed at avoiding overlapping patches. For a complex protein including about 2000 atoms, which generates more that 1000 candidate patches and nearly 300 000 constraints, this optimization problem can be solved in a couple of minutes by a software package for linear programming problems, such as ILOG CPLEX, with Mixed Integer Programming option.

4. Experimental results

The tests we carried out are aimed at evaluating the capability of the algorithm to generate clusterings that effectively capture protein similarities. We carried out an indirect evaluation in the context of the framework for protein classification described in Section 2, where clustering is functional to surface comparison and to the mining of frequent surface patterns.

We performed our tests on a set of 25 real proteins, already used to test surface based classification in [7], and belonging to five families: hemoglobins (1a00, 1a01, 1a0u, 1a0v, 1a0w, 1a0x, 1a0z, 1gzx), ureases (1a5k, 1a5l, 1a5m, 1a5n, 1a5o), crambin-like (1jxt, 1jxu, 1jxw, 1jxx, 1jxy), seryl-tRNA synthetases (1ser, 1ses, 1set, 1sry) and hydrolases (1tca, 1tcb, 1tcc). All tests are based on the following setting for parameters: α_{cur}=0.035, β_{cur}=0.03, α_{elp}=0.275, β_{elp}=0.05, σ=0.75, r_{max} = 12.

Figure 1 shows, in its top section, an example of clustering. The left part shows the surface of protein 1a00; blue, red, and white respectively mean positive, negative, and neutral electrostatic potential. The right part shows the clustering obtained by emphasizing

patches with different colors; the portions of surface in black are not assigned to any patch.

As to classification, for each couple of proteins p_i, p_j we first compute a similarity index $Sim(p_i, p_j)$ according to the number of patterns shared between p_i and p_j [1]. In Figure 3 we show the all-for-all similarities between proteins (black means $Sim = 0$, white $Sim = 1$; a logarithmic scale is used). The five families are very well recognizable as five light squares, which means that intra-family similarity is much higher than inter-family similarity.

Finally, the resulting dendrogram is depicted in Figure 4. Remarkably, the five families are correctly separated, thus confirming the effectiveness of our approach in correctly capturing surface similarities.

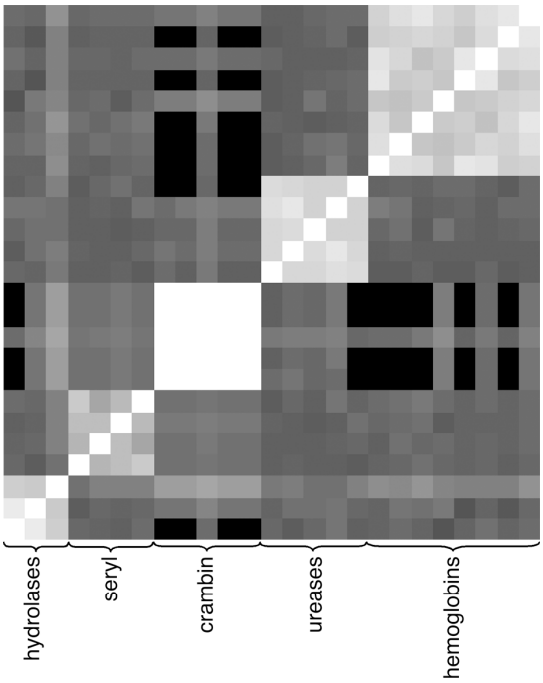

Figure 3. Surface similarities

5. Discussion

In this paper we proposed a clustering technique for protein surfaces aimed at creating homogeneous and regular surface patches for protein classification purposes. A comparison with the results obtained in [7] yields that, while a correct classification is obtained by both approaches, our surface representation turns out to be more compact since it determines a lower number of patches, each carrying a more comprehensive information. In fact the approach in [7], differently from ours, does not allow the use of biochemical information due to its high computational demand.

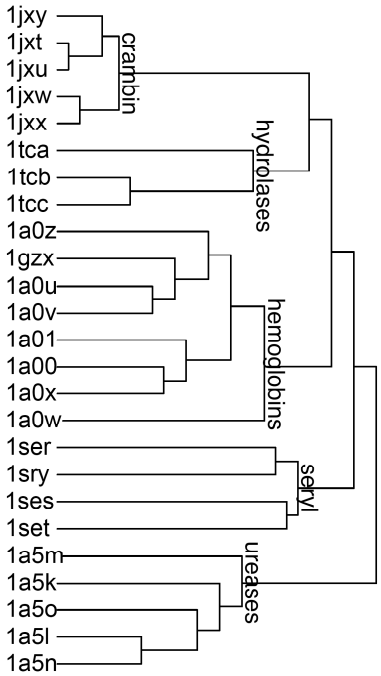

Figure 4. Classification dendrogram

6. References

[1] Baldacci L. and M. Golfarelli, "Mining complex patterns from molecular surfaces", in *Proc. Int. Work. on Biological Data Management*, 2005.

[2] Connolly M., "Solvent-accessible surface of proteins and nucleic acids", *Science*, 221:709–713, 1983.

[3] Fischer D., H. Wolfson, S. Lin, and R. Nussinov, "Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities", *Protein Sci.*, 3:769–778, 1994.

[4] Kauvar L. and H. Villar, "Deciphering cryptic similarities in protein binding sites", *Curr. Opin. Biotechnol.*, 9(4):390–394, 1998.

[5] Koradi R., M. Billeter, and K. Wuthrich, "MOLMOL: a program for display and analysis of macromolecular structures", *Molecular Graphics*, 14:51–55, 1996.

[6] Kyte J. and R. Doolittle, "A simple method for displaying the hydropathic character of a protein", *Molecular Biology*, 157(1):105–132, 1982.

[7] Lozano M. A. and F. Escolano, "Protein classification by matching and clustering surface graphs", *Pattern Recognition*, 2006 (to appear).

[8] Pauly M., M. Gross, and L. P. Kobbelt, "Efficient Simplification of Point-Sampled Surfaces", in *Proc. 13th IEEE Visualization Conf. (VIS)*, 2002, pp. 163-170.

[9] Pickering, S., A. Bulpitt, N. Efford, N. Gold, and D. Westhead, "AI-based algorithms for protein surface comparisons", *Comp. and Chem.,* 26:79–84, 2001.

[10] Wang X., "Finding patterns on protein surfaces: Algorithms and applications to protein classification", *IEEE Trans. Knowl. Data Eng.*, 17(8):1065–1078, 2005.