31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Structural Feature Extraction Protocol for Classifying Reversible Membrane Binding Protein Domains.

Morten Källberg and Hui Lu

Abstract— **Machine learning based classification protocols for automated function annotation of protein structures have in many instances proven superior to simpler sequence based procedures. Here we present an automated method for extracting features from protein structures by construction of surface patches to be used in such protocols. The utility of the developed patch-growing procedure is exemplified by its ability to identify reversible membrane binding domains from the C1, C2, and PH families.**

I. INTRODUCTION

Complex signaling networks involving both protein-protein and protein-lipid interactions allow for the rapid synchronization between cell activity and external environment. One vehicle for such processes is the reversible translocation of cytoplasmic proteins to cellular membranes [1]: By increasing the effective concentration of two interaction partners in a confined space close to the membrane a signal exchange becomes much more likely. Members of a number of domains families (such as C1, C2, PH, FYVE, ENTH and PX domains) have been found to drive the association with membranes by means of a collection of common mechanisms. In particular, properties such as the nonspecific electrostatic attraction between anionic membranes and cationic surface residues [2], association of hydrophobic surface residues with the membrane hydrocarbon core [3], and the specific interaction between key residues and lipid head-groups through hydrogen-bonding [4] have been found to be of major importance (albeit not all mechanisms are equally prominent in every families).

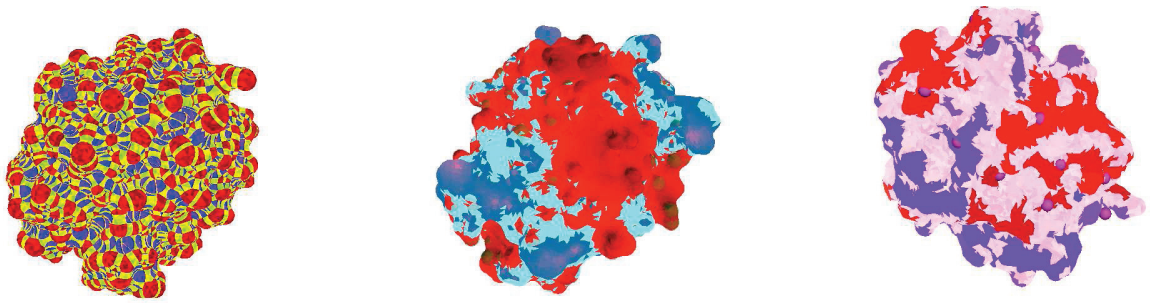
To identify and predict which protein domains can bind to membrane is of great importance for understanding cell signaling networks. Simple procedures based on sequence similarity for function annotation do, however, not give the desired accuracy in uncovering such mechanisms. Therefore we aim at developing more advanced machine learning based approaches for this task. In past decades a vast collection of well-performing machine learning algorithms have been developed [5], [6]. The concern when constructing a classification procedure is thus how to encode the information from the function of each annotated domain in the trainingset into a numerical vector capturing its essential properties.

In this work we present an approach for automating the feature calculation task for protein domains with known structure. The characteristic properties of the structure is captured by identifying continuous regions of the solvent exposed surface, so-called surface patches, defined by physical or chemical quantities (electrostatic potential, hydrophobicity etc.) common to the this specific area. We illustrate the method by calculating features believed to be of importance in identifying protein domains involved in reversible binding to plasmamembranes, it should, however, be noted that the technique as such is more general and can easily be tailored to address any type of structure-based classification scenario. This in contrast to previous works developing patch-centered surface representations where either local structure similarity was used [7] to screen a library of functional sites, or unsupervised clustering-based tecniques were applied [8], [9].

II. AUTOMATED FEATURE EXTRACTION THROUGH PATCH GROWING

The steps of patch growing detailed below are outlined in Fig. (1). The basic idea is as follows:

Department of Bioengineering - Bioinformatics program, University of Illinois at Chicago, 851 S. Morgan St. Chicago, IL 60607-7052 huilu@uic.edu



(a) The solvent-excluded surface as calculated by the MSMS (Maximal Speed Molecular Surface) algorithm, the coloring of the triangles representing the surface represent the number of probe contact made (see [10] for details).

(b) Mapping of the electrostatic potential onto the surface, dark-red regions indicate higly negative potential values while dark-blue regions indicate highly positive potential values.

(c) The patches grown on the surface with parameter $C = 60$, red regions indicate patches with negative potential value, purple regions indicate patches with positive surface value, and pink regions indicate location with mixed potential values.

Fig. 1. The three steps in determining surface patches for a certain quantity for protein structure PDB-id 1a53, here illustrated with the electrostatic potential of the structure.

First, the surface is defined as a collection of neighboring triangles (Fig. 1(a)), second, a numerical representation of the quantity of interest is associated with each triangle Fig. 1(b), and finally the patches that are most highly correlated with the function of the structure are defined (Fig. 1(c)).

A. Surface patch definition

By using the definition of solvent-excluded surface (SES) in [10] the topological boundary defined by the Van der Waals radius of the atoms in the structure of interest is determined by use of the MSMS algorithm developed by Sanner [10]. The final SES is expressed by a triangulation procedure and thus results in a collection of neighboring triangles representing the molecular surface.

For now lets assume that each triangle on the surface is associated with a numerical value corresponding to the quantity that forms the basis for patch growing. We will denote this value for a triangle t by $t.val$ and the distance between the centroids of triangles t_1 and t_2 by $dist(t_1, t_2)$. Furthermore, $t.neigh$ will denote the neighbor triangles of t, meaning those that share an edge with t, and $t.included$ will be a boolean flag indicating whether a given triangle has been included in a patch. The collection of patches is then found by repeating the following recursive procedure until all surface triangle have been included in a patch: Choose a random triangle that has not yet been included in a patch and extend the patch from here according to the procedure outlined in Fig. 2, repeat until all triangles have been included in a patch.

```
GROW-PATCH(Seed triangle T):
 for t1 in T.neigh:
   if NOT t1.included AND
   |t1.val - T.val|/dist(T,t1)<C:
       Add t1 to the current patch
       t1.included = TRUE
       GROW-PATCH(t1)
```

Fig. 2. Pseudocode for the patch growing procedure. C is a constant determining whether the patch should be extended to a given neighbor triangle.

The constant C in the `GROW-PATCH`-method (Fig. 2) is used to determine if a patch should be extended in a given direction. An appropriate value for C needs to be set for each patch type of interest. A C-value can of course be determined manually by simple visual inspection of the patches in the molecular surface, however, a fully automate procedure is desirable. Since our aim is to use the patch-growing in context of machine-learning, determining the C-value can be made part of the learning procedure. As we will have a annotated trainingset available the final C-value is determined by its ability to grow patches that separate the two

groups in the trainingset, as measured by the Fisher-score [11].

B. Mapping quantities onto the surface

In order to do the patch-growing we need to assign values from the quantity of interest to each triangle on the SES. Here we give examples of how this can be done for both spatial and residue/atom based data.

(1) The electrostatic potential of a structure can be calculated by solving the Poisson-Boltzmann (PB) equation numerically using a finite difference scheme as implemented in APBS [12]. The spatial potential values are mapped onto the surface by taking a weighted average of the 8 discrete data points closest to the point 1 A from the triangle surface in the direction of its normal vector. (2) Hydrophobicity values are assigned to the surface based on the Kyte-Doolittle value of the amino acid that gave rise to the triangle of interest [13]. (3) Hydrogen-bonding is mapped to the surface by determining if an atom is capable of forming a hydrogen bond, indicated by setting $t.val = 1$.

III. EXPERIMENTS WITH C1, C2, AND PH DOMAINS

For testing the value of the patch growing procedure in identifying structures of a particular function, a collection of domains from families known to be involved in reversible membrane targeting was collected. These were annotated as either binding membrane or as having other function (for instance participating in protein-protein interaction)

A. Datasets

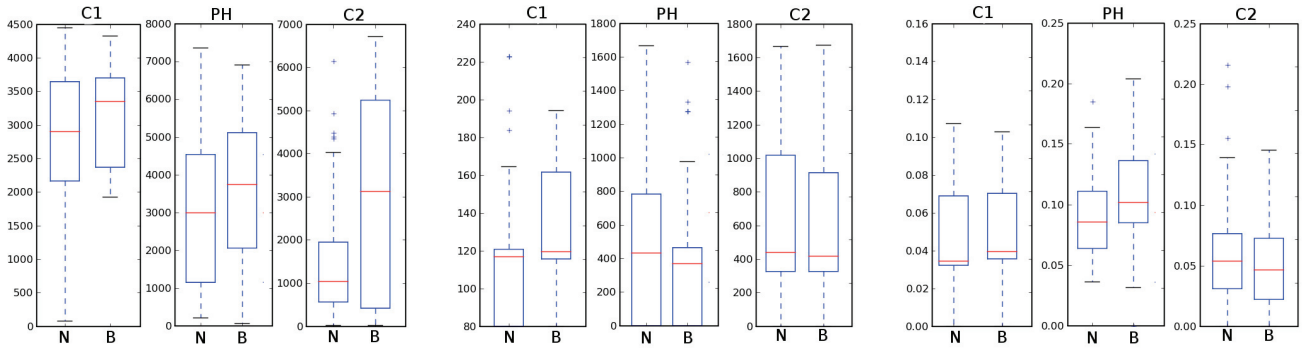
The dataset was constructed from our online resource for Membrane Targeting Domains (MeTaDor) [14]. A subset of domains that do not share more than 75% sequence similarity and with known binding properties were extracted. For the domains in the set not having experimental structures available homology models were created (domains for which no template with at least 40% sequence similarity existed were discarded). The resulting datasets have binding/non-binding counts of 33/22, 63/27, and 70/88 for C1, C2, and PH domains, respectively.

B. Separation by electrostatic potential, hydrophobicity, and hydrogen-bonding capability

The above framework was used to grow patches based on the three quantities outlined in Section II-B. As can be seen in Fig. 3 the features derived from these patches do in general show good discriminatory power between the two classes of domains, although some features prove more instrumental in specific families. First, we compare the cumulative area of the five largest patches with positive potential value grown based on the electrostatic calculations. It is evident from Fig. 3(a) that membrane binding domains do in general have a larger positive surface area than non-binding, regardless of family. This observation correlates well with accepted models proposed for membrane-targeting domains, which suggest that initial non-specific translocation to membranes is usually due to attraction between positively charged protein domains and negatively charged lipid-head groups [4].

The next feature we inspect is the size of the largest hydrophobic surface patch. As can be seen in Fig. 3(b) this features seems to discriminate well amongst binding and non-binding C1 domains, but has near identical distribution for C2 and PH domains. Again there is good correlation with known properties of the binding mechanisms for these domains. Many C1 domains strengthen the binding to membranes by insertion of a collection of hydrophobic residues clustered on one side of the structure, into the hydrocarbon core of the membrane [15], whereas C2 and PH domains most often do not penetrate deeply into the membrane.

Finally, we compare the hydrogen-bonding capability of the binding and non-binding domains by calculating the percentage of the SES covered by hydrogen-bonding patches. Fig. 3(c) shows that this feature mainly has discriminatory power between binding and non-binding cases in the PH domain family, while there is little difference in the cases from C1 and C2 domain families. Though all three domains families form hydrogen-bonds with lipids upon binding membrane, the effect is most prominent in PH domains which can have as many as three PIP_3 binding-sites, all forming



(a) Cumulative area of the five largest surface patches with positive potential value compared among the 3 families.

(b) The area of the largest hydrophobic patch compared among the 3 families.

(c) The percentage of surface area covered by hydrogen-bonding patches compared among the 3 families.

Fig. 3. The comparison of 3 features obtained from patches grown with 3 different quantities mapped onto the solvent-excluded surface. For each feature the distribution among Binding (B) and Non-binding (N) domains for C1, PH, and C2 domains are compared.

several hydrogen-bonds [16].

IV. CONCLUSION

We have constructed a framework for feature generation from protein structures to be utilized in machine learning application of protein function annotation. As illustrated with the reversible membrane targeting domain datasets, features derived in this fashion show good discriminatory power in separating binding and non-binding domains making construction of strong machine learning classifiers plausible. Furthermore, the framework can easily be extended to handle other structure based classification tasks.

REFERENCES

[1] M. N. Teruel and T. Meyer, "Translocation and reversible localization of signaling proteins: a dynamic future for signal transduction." *Cell*, vol. 103, no. 2, pp. 181–184, Oct 2000.

[2] A. Mulgrew-Nesbitt, K. Diraviyam, J. Wang, S. Singh, P. Murray, Z. Li, L. Rogers, N. Mirkovic, and D. Murraycons, "The role of electrostatics in protein-membrane interactions." *Biochim Biophys Acta*, vol. 1761, no. 8, pp. 812–826, Aug 2006.

[3] R. V. Stahelin and et. al, "Contrasting membrane interaction mechanisms of ap180 n-terminal homology (anth) and epsin n-terminal homology (enth) domains." *J Biol Chem*, vol. 278, no. 31, pp. 28 993–28 999, Aug 2003.

[4] W. Cho and R. V. Stahelin, "Membrane-protein interactions in cell signaling and membrane trafficking." *Annu Rev Biophys Biomol Struct*, vol. 34, pp. 119–151, 2005.

[5] C. Cortes and V. Vapnik, "Support-vector networks." *Machine Learning*, vol. 20, pp. 273–97, 1995.

[6] J. R. Quinlan, "Induction of decission trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[7] F. Ferr, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich, "Surface: a database of protein surface regions for functional annotation." *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D240–D244, Jan 2004.

[8] L. Baldacci, M. Golfarelli, A. Lumini, and S. Rizzi, "Clustering techniques for protein surfaces," *Pattern Recognition*, vol. 39, p. 2370 2382, 2006.

[9] M. Lozano and F. Escolano, "Protein classification by matching and clustering surface graphs," *Pattern Recognition*, vol. 39 (4), pp. 499–736, 2006.

[10] M. F. Sanner, A. J. Olson, and J. C. Spehner, "Reduced surface: an efficient way to compute molecular surfaces." *Biopolymers*, vol. 38, no. 3, pp. 305–320, Mar 1996.

[11] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *In Advances in Neural Information Processing Systems 11*. MIT Press, 1999, pp. 487–493.

[12] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: application to microtubules and the ribosome." *Proc Natl Acad Sci U S A*, vol. 98, no. 18, pp. 10 037–10 041, Aug 2001.

[13] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein." *J Mol Biol*, vol. 157, no. 1, pp. 105–132, May 1982.

[14] N. Bhardwaj, R. V. Stahelin, G. Zhao, W. Cho, and H. Lu, "Metador: a comprehensive resource for membrane targeting domains and their host proteins." *Bioinformatics*, vol. 23, no. 22, pp. 3110–3112, Nov 2007.

[15] W. Cho, "Membrane targeting by c1 and c2 domains." *J Biol Chem*, vol. 276, no. 35, pp. 32 407–32 410, Aug 2001.

[16] J. H. Hurley and S. Misra, "Signaling and subcellular targeting by membrane-binding domains." *Annu Rev Biophys Biomol Struct*, vol. 29, pp. 49–79, 2000.