Protein Surface Atom Neighborhoods Functional Imaging

Paul Dan Cristea, Octavian Arsene, Rodica Tuduce
University "Politehnica" of Bucharest
Romania
pcristea@dsp.pub.ro

Dan Nicolau
Liverpool University
United Kingdom
d.nicolau@liverpool.ac.uk

***Abstract*—A protein functional imaging in terms of their surface hydrophobicity distribution is presented. Atom level hydrophobicities, obtained from amino acid hydrophobicities, are used to compute approximately equivalent hydrophobicity density distributions around each surface atom, in a standardized octagonal framework. Surface atom neighborhoods are classified in terms of their resemblance or based on a vector description.**

Keywords- protein structure; protein surface; atom hydrophobicity; surface hydrophobicity; resemblance; clustering.

I. INTRODUCTION

We have analyzed a set of 36 molecules from the Protein Database Bank (PDB) [1], computing the hydrophobicity distribution on their surface, as requested by the EU research project BISNES (Bio-Inspired Self-assembled Nano-Enabled Surfaces) [2]. The solvent-excluded surface of the molecules has been determined with the Connolly algorithm [3,4] and atom-level hydrophobicities, consistent with amino acid hydrophobicities [5,6], have been used to further describe the surface hydrophobicity distribution [7]. To simplify the comparison of hydrophobicity distributions on protein surfaces, we have introduced an approximately equivalent description of surface atom neighborhoods in terms of hydrophobicity densities in a standardized octagonal frame. [8,9].

Each neighborhood frame was divided into equal area annular sector patches, as illustrated in Fig. 1. The density of a chosen property (e.g., the hydrophobicity), is computed cumulatively, by the simple addition of the discrete values for all atoms in the corresponding patch. This approach generates a pattern which describes functionally the molecule, in the proximity of each surface atom *A*, from the point of view of the considered property [10].

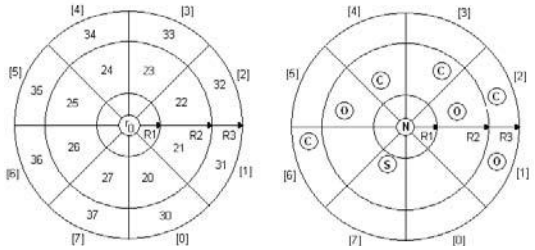

Figure 1. Octagonal standardized pattern around a surface atom and example of neighboring atoms distribution

II. HYDROPHOBICITY SURFACE DENSITIES

The neighborhood of a surface atom A consists of a central circle with radius R_1 (patch $P_A(1)$) and area $A_1 = \pi R_1^2$, and of $8(n_{\max} -1)$ patches in the shape of annular sectors, obtained by the partition of each annuli between the circles of radii R_{n-1} and R_n in eight equal patches $P_A(n, k)$, $n \in \{2,\ldots, n_{\max}\}$, $k \in \{0,\ldots,7\}$) of area $A_{nk} = \pi(R_n^2 - R_{n-1}^2)/8$ [8]. For most proteins, a surface atom neighborhood comprises $n_{\max} = 8$ or 9 circles, up to the radius $R_{n_{\max}}$ for which the distance between the atoms in the considered neighborhood and the atoms on a flat surface tangent to the central atom exceeds h=2 Å – the maximum distance at which atoms can still interact by hydrogen bonds. For small molecules, the outermost patches can remain empty, without neighboring atoms. Depending on the specific shape of a molecule, some other patches can also be empty. Because we imposed that all patches have equal area, the radiuses of the successive annuli are in the sequence:

$$R_n = R_1 \cdot \sqrt{8(n-1)+1}\,. \tag{1}$$

The density of hydrophobicity, in arbitrary units, can be computed by the addition of the atom hydrophobicities H_h:

$$H_A(1) = \sum_{h \in P_A(1)} H_h, \quad H_A(n,k) = \sum_{h \in P_A(n,k)} H_h, \tag{2}$$

in the central patch $P_A(1)$, and in the annuli sectors between the circles of radii R_{n-1} and R_n, $n \in \{1, .., n_{\max}\}$ and the angular sector $k \in \{0, .., 7\}$, respectively. The atomic hydrophobicities de-convoluted from the amino-acid hydrophobicities, corrected for the percentage of the atom surface included in the Connolly surface, have been used [3,4].

III. RESEMBLANCE OF SURFACE ATOM NEIGHBORHOODS

The *resemblance* of two surface atom neighborhoods, rotated by h steps with respect to each other, is defined [7]as:

$$R(A,B,h) = H_A(1)\, H_B(1) + \sum_{\substack{n \in \{2,\ldots,n_{\max}\} \\ k \in \{0,\ldots,7\}}} H_A(n,k)\, H_B(n, k \oplus h); \quad h \in \{0,\ldots,7\}, \tag{3}$$

where $H_A(1)$, $H_B(1)$ are the total hydrophobicities of the central patches of radius R_1, $H_A(n, k)$, $H_B(n, k)$ – the hydrophobicities of the annuli sectors rotated around the common normal with h

steps with respect to each other, n_{max} – the maximum circle rank (the resolution). The symbol $\oplus$ designates the sum modulo 7.

The resemblance varies when the two neighborhoods are rotated around the common normal. For each pair of surface atoms A and B, the maximum and the minimum resemblance are defined by:

$$R_{\max}(A,B) = \max_{h\in\{0,\ldots,a-1\}} R(A,B,h) \qquad (5)$$

$$R_{\min}(A,B) = \min_{h\in\{0,\ldots,a-1\}} R(A,B,h) \quad .$$

Because two interacting molecules can usually move with respect to each other, these magnitudes are more significant when comparing or when estimating the interaction of two surface atom neighborhoods.

The *similitude* and the *interaction* of a pair of atom neighborhoods are defined as their *resemblance* for parallel and, respectively, anti-parallel orientations of the vectors normal on the molecular surfaces in the superposed contact points *A=B*, as shown in Fig. 2.

For illustration, Fig. 3 shows the histogram of the maximum interactions for all pairs of surface atoms of the proteins with the labels 135L and 1HZH in PDB. All mutual orientations of the two surfaces are considered for each pair of surface atoms and the largest value of the interaction is retained in each case.

Similarly, Fig. 4 shows the histogram of the minimum interactions for all pairs of surface atoms of the same proteins 135L and 1HZH. The regularity of these histograms indicates a limited variety of the surface atom neighborhoods.

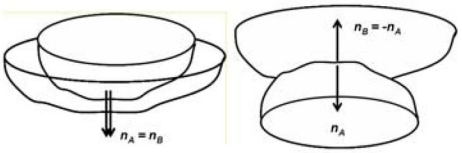

Figure 2. Parallel and anti-parallel orientation of vectors normal on the molecular surfaces in the superposed contact points A=B.

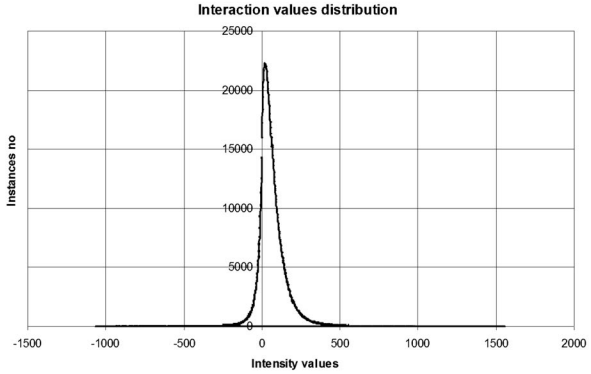

Figure 3. Maximum interaction histograms for all pairs of surface atoms of the proteins 135L and 1HZH.

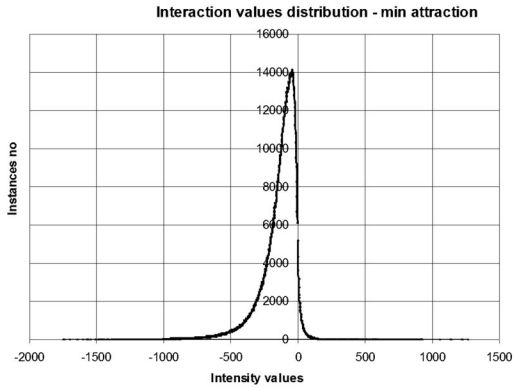

Figure 4. Minimum interaction histograms for all pairs of surface atoms of the proteins 135L and 1HZH.

IV. Vector Classification of Surface Atom Neighborhoods

A functional image oriented representation of protein surfaces and the corresponding software tool have been developed to interactively explore the hydrophobicity distribution on the molecular surface. The surface atom neighborhoods have been described by vectors of 65 components which specify the hydrophobicity densities (2) in each patch of the standardized octagonal frame defined in Section II. The comparison of the neighborhoods has been performed both globally, in termes of *similitude* and *interaction,* and by their clustering using the vector description.

Because the relative angular position of two surface atom neighborhoods can vary arbitrarily, it is necessary to pre-process the vectors describing the neighborhoods, to bring them in similar positions before applying any classification in algorithm. The calibration pattern shown in Fig. 5 has been used for this purpose. The pattern has patches of clockwise decreasing hydrophobicities, starting from the highest hydrophobicity density (3.89) in the sector 1, passing through zero hydrophobicity density in sector 6, and reaching the lowest value (-1.007) in sector 8 (7 and 8 are thus hydrophilic). Each surface atom neighborhood is rotated to a position in which its resemblance to the calibration pattern is maximized.

The WEKA (Waikato Environment for Knowledge Analysis) API was used as clustering engine. Fig. 6. shows the results of the clustering in terms of the *Average Hydrophobicity vs. Average Resemblance* dependence.

Figure 5. Surface atom neighborhoods calibration pattern with clockwise decreasing hydrophobicities.

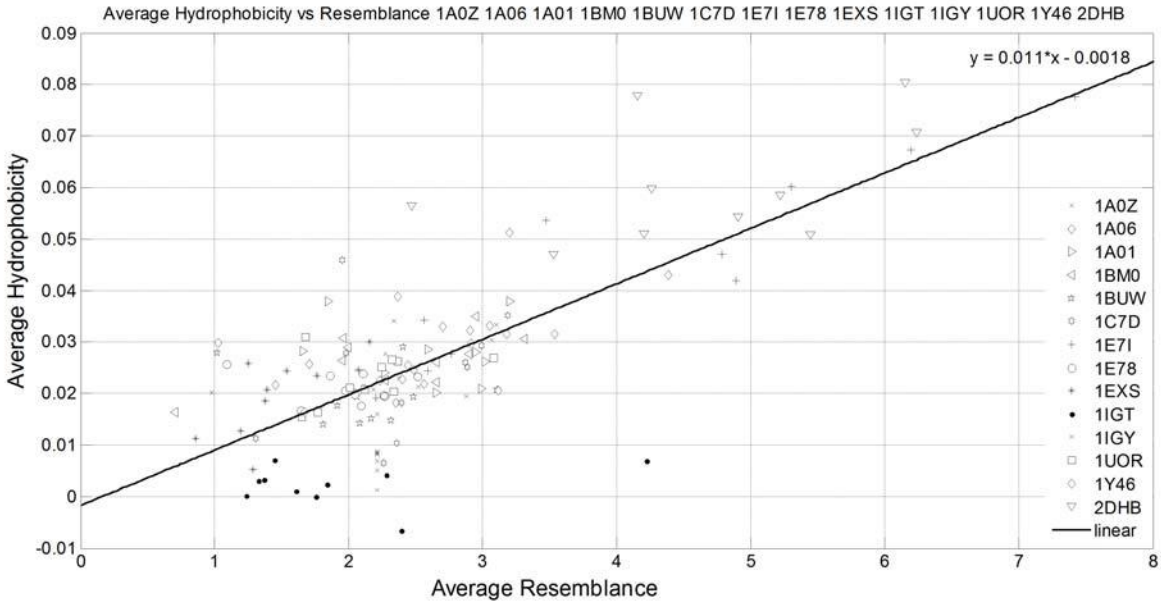

Figure 6. Clustering of surface atom neighborhoods for the 1A0Z, 1A06, 1A01, 1BM0, 1BUW, 1C7D, 1E7I, 1E78,1EXS, 1IGT,1GY,1UOR, 1Y46, 2DHB proteins [], expressed by the dependence of the *Average Hydrophobicity* on the *Average Resemblance* of clusters to a reference pattern, after re-orienting individual surface atom neiborhoods to maximize their resemblance to the reference pattern.

The *Average Hydrophobicity* of the surface atom neighborhood clusters generated by WEKA is given as a function of their *Average* (maximum) *Resemblance* to the reference pattern in Fig. 5, after the re-orientation of the individual pattern in each cluster. Data for the larger molecules (1A0Z, 1A06, 1A01, 1BM0, 1BUW, 1C7D, 1E7I, 1E78,1EXS, 1IGT,1GY,1UOR, 1Y46, 2DHB) in the set of 36 studied proteins have been used in Fig. 6. An approximately linear dependence has been found for the 140 clusters, with a larger spreading of data for the small values of average resemblance. Notice that the resemblance to the reference pattern is maximized for each individual neighborhood in a cluster, whereas the average resemblance of for all the patterns is given for every cluster.

Results of clustering at the vector level are given in Figs. 7-10, for the 1E71, 1Y46, 1UOR and 2DHB protein surface atom neighborhoods.

Each neighborhood is described by a hydrophobicity density vector with 65 components, corresponding to the central circle and to the 8x8 annular sectors. The goal of the analysis is to find patterns on the protein surface neighborhoods in terms of hydrophobicity. Because there are no priorly defined classes of hydrophobicity densities, an unsupervised learning scenario has to be used. Clustering is used to group items that seem to fall naturally together. The output takes the form of a list specifying how the instances fall into clusters. The success of clustering is often measured subjectively in terms of how useful the result appears to a human user.

The basic clustering technique is the *k*-means. The user specifies in advance how many clusters are being sought, the *k* parameter. Then, *k* points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metrics.

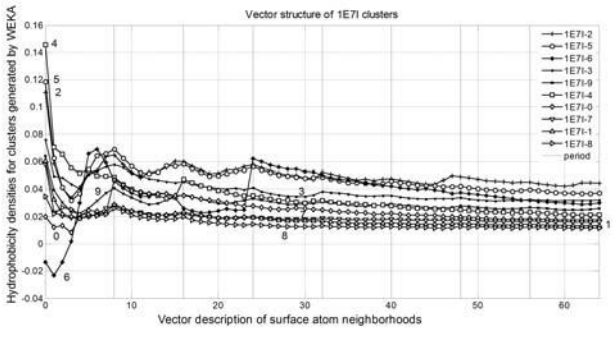

Figure 7. Vector clustering of 1E71 protein surface atom neighborhoods.

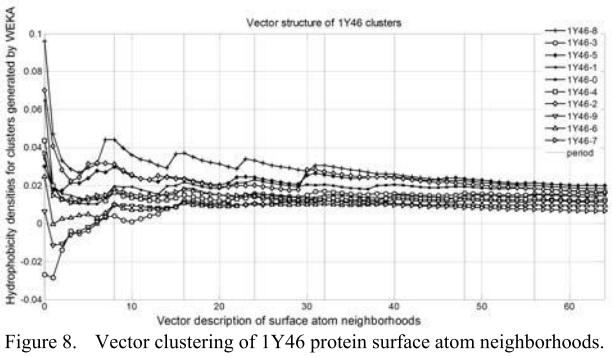

Figure 8. Vector clustering of 1Y46 protein surface atom neighborhoods.

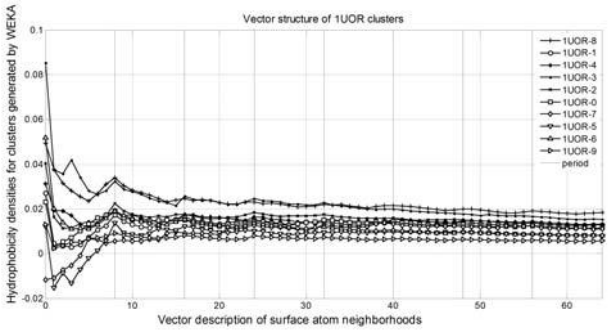

Figure 9. Vector clustering of 1UOR protein surface atom neighborhoods.

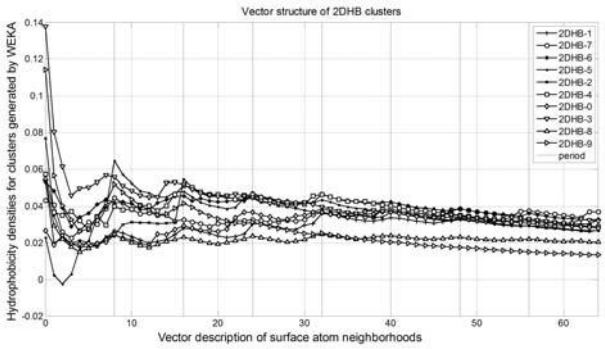

Figure 10. Vector clustering of 2DHB protein surface atom neighborhoods.

We have used the expectation-maximization algorithm for clustering the hydrophobicity densities of the atoms vicinities, one of the most powerful methods for finding a maximum likelihood solution for models with hidden variables [11,12].

The WEKA EM implementation requires the input parameters:

- Number of clusters to generate, chosen 10 based on the results obtained from tests with 5 to 25 clusters;
- Maximum number of iterations, chosen to 100 iterations;

Maximum allowable standard deviation for the density calculation, chosen 10^{-6}.

The figures give the average hydrophobicity densities for the 65 components describing each of the ten clusters in which are classified similar patterns. Prototype images have been constructed for each of the clusters by using their average hydrophobicity vectors. These images, not shown here to avoid the need of colors, represent a set of "physiognomies" corresponding to the pattern at the surface of each studied protein.

V. CONCLUSIONS

We continued the study of protein surfaces has with the classification of local molecule properties using resemblance and vector descriptions [10]. The global values of these magnitudes, the histograms of their distribution for all the surface atoms, and the actual structure of the surface atom neighborhoods are taken into account.

The classification allows predicting the behavior of protein molecules when interacting with each other, or with a nanostructured surface. Further work will include not only the study of the interactions determined by the hydrophobicities, de-convoluted at the level of atoms [6], but also the effects of the electrical interactions and the way these interactions are influenced by the pH. The cumulated effect of the two types of interactions will be expressed in a coherent way, allowing to compute the resulting similitude and interaction of two surface atom neighborhoods. This approach will result in a better description of the complex phenomena involved in the protein multi-parameter interactions. A web approach is also considered, to facilitate the access of academia and industry to the new results.

ACKNOWLEDGMENT

The work was partially supported by the project 214538 – 2008 - BISNES – "Bio-Inspired Self-assembled Nano-Enabled Surfaces," in the framework of the NMP-2007-1.1-2 Self-assembling and self-organization, NMP-2007-1.1-1 Nano-scale mechanisms of bio/non-bio interactions.

REFERENCES

[1] Protein Data Bank [Online], http://www.rcsb.org/pdb/home/home.doc.

[2] Project 214538 – 2008 - BISNES – "Bio-Inspired Self-assembled Nano-Enabled Surfaces," http://www.bisnes4eu.com/

[3] M. L. Connolly, "MS: Molecular Surface Program," *QCPE Program 429*, Quantum Chemistry Program Exchange, Univ. of Indiana, Bloomington, 1983.

[4] M. L. Connolly, "Molecular Surfaces: A Review," Network Science (On line), vol.2(4), 1996, http://www.awod.com/netsci/Science/Compchem/feature14.html.

[5] P. A. Karplus, "Hydrophobicity regained," *Protein Science*, vol. 6, pp. 1302-1307, 1997.

[6] M. Held and D. V. Nicolau, (2007), "Estimation of atomic hydrophobicities using molecular dynamics simulation of peptides," Proc. of SPIE, 6799, Modelling and THZ Technology, 6799-16, 1-7.

[7] D. V. Nicolau, F. Fulga, and D. V. Nicolau, (2003), "A new program to compute the surface properties of biomolecules," Asia-Pacific Biotech, 7(3), 29-34.

[8] P. D. Cristea, Rodica Tuduce, O. Arsene, D. V. Nicolau, F. Fulga, "Multi-threading Protein Surface Functional Description," NEUREL 2010, Belgrade, Serbia, September 23-25, 2010, Proc. of NEUREL 2010, pp. 1075-1078.

[9] P. D. Cristea, Rodica Tuduce, O. Arsene, Alina Dinca, D. V. Nicolau and F. Fulga, "Modeling of Biological Nanostructured Surfaces," Proc. of SPIE, 7574, Nanoscale Imaging, Sensing, and Actuation for Biomedical Applications VII, 2010.

[10] P. D. Cristea, Rodica Tuduce, O. Arsene, D. V. Nicolau, "Functional Nanoscale Imaging of Protein Surfaces," *BiOS SPIE Photonics West – Nanoscale Imaging, Sensing, and Actuation for Biomedical Applications Conference*, San Francisco, California, USA, 2011.

[11] Russell, S., Norvig, P., "Artificial Intelligence. A Modern Approach," Prentice Hall, 3rd edition, 2010.

[12] Bishop, C., "Pattern Recognition and Machine Learning," Springer, 2006.