Localization of ligand binding site in proteins identified in silico

Michal Brylinski · Marek Kochanczyk · Elzbieta Broniatowska · Irena Roterman

Received: 20 November 2006 / Accepted: 26 February 2007 / Published online: 30 March 2007
© Springer-Verlag 2007

Abstract Knowledge-based models for protein folding assume that the early-stage structural form of a polypeptide is determined by the backbone conformation, followed by hydrophobic collapse. Side chain–side chain interactions, mostly of hydrophobic character, lead to the formation of the hydrophobic core, which seems to stabilize the structure of the protein in its natural environment. The fuzzy-oil-drop model is employed to represent the idealized hydrophobicity distribution in the protein molecule. Comparing it with the one empirically observed in the protein molecule reveals that they are not in agreement. It is shown in this study that the irregularity of hydrophobic distributions is aim-oriented. The character and strength of these irregularities in the organization of the hydrophobic core point to the specificity of a particular protein’s structure/function. When the location of these irregularities is determined versus the idealized fuzzy-oil-drop, function-related areas in the protein molecule can be identified. The presented model can also be used to identify ways in which protein–protein complexes can possibly be created. Active sites can be predicted for any protein structure according to the presented model with the free prediction server at http://www.bioinformatics.cm-uj.krakow.pl/activesite. The implication based on the model presented in this work suggests the necessity of active presence of ligand during the protein folding process simulation.

Keywords Hydrophobic collapse · Protein folding · Active site · Ligand binding

Introduction

Since the classic work by Kauzmann [1], hydrophobic interactions have been confirmed to play a crucial role in forming and stabilizing the protein tertiary structure [2–5]. It is generally accepted that globular proteins consist of a hydrophobic core and a hydrophilic surface [6, 7]. The way in which the amino acid sequence partitions a protein into its interior and exterior has been described [8]. The inside regions are densely packed chain sites where hydrophobicity is observed to be at a local maximum, whereas the outside regions correspond to less densely packed sites where hydrophobicity is at a local minimum. The spatial distribution of amino acid hydrophobicity has been used to differentiate native and non-native protein structures [9–12]. Irbäck and co-workers brought forward a proof for nonrandom hydrophobicity structures in protein chains [13]. A second-order hydrophobic moment was discussed for description of protein hydrophobicity [14]. Detailed analyses of the spatial variation of hydrophobicity, focused on the region of transition between the protein interior and exterior, were carried out for 30 relatively diverse globular
proteins as well as for 14 decoys [15]. Apart from soluble proteins, the distribution of apolar and polar residues has provided comprehensive information about transmembrane protein architecture [16–19]. The hydrophobic effect has been suggested to be the dominant driving force in protein folding [20–22].

The model presented in this paper is oriented on localization of the area responsible for ligand binding or protein–protein complex creation, based on the characteristics of the spatial distribution of hydrophobicity in a protein molecule. It is assumed that hydrophobicity changes from the protein interior (maximal hydrophobicity) to the exterior (close to zero hydrophobicity) according to the three-dimensional Gaussian distribution. It is generally accepted that the core region is not well described by a spheroid of buried residues surrounded by surface residues due to hydrophobic channels that permeate the molecule [23, 24]. This being so, we should be able to identify regions with high deviation versus the ideal model by making a simple comparison of the theoretical (idealized according to the Gaussian function) and empirical spatial distribution of hydrophobicity in a protein. The regions recognized by high hydrophobicity density differences seem to reveal functionally important sites in proteins.

The model presented here can be used for structure-based prediction of the localization of active sites in proteins of unknown function. It can also be used for qualitative and quantitative analysis of known protein–protein or protein–ligand complexes. The statistical method introduced for assignment of theoretical hydrophobicity to small ligands, combined with searches for equivalent hydrophobic cavities, may also be very useful for docking simulations.

Materials and methods

Data

The following structures were selected for analysis of proteins complexed with small ligands: myoglobin (PDB ID: 1A6M) [25], subtilisin DY (PDB ID: 1BH6) [26], carboxypeptidase A2 (PDB ID: 1DTD) [27], chymotrypsin (PDB ID: 1GG6) [28], c-type lysozyme (PDB ID: 1LMQ) [29] and ribonuclease (PDB ID: 1RGE) [30]. The following CAPRI targets were selected for studying the structures of protein–protein complexes: transcriptional antiterminator protein LicT [target 09, homodimer, PDB ID: 1H99] [31], cohesin-dockerin complex [targets 11 and 12, PDB ID: 1OHZ] [32] and the complex between protein serine/threonine phosphatase-1 (delta) and N-terminal domain of the myosin phosphatase targeting subunit MYPT1 [target 14, PDB ID: 1S70] [33].

Theoretical (expected) hydrophobic core

The fuzzy-oil-drop model, which was applied to simulate so-called hydrophobic collapse in the protein folding process, represents the theoretical, idealized hydrophobic core of the protein molecule. Such a hydrophobic core surrounded by the spheres of gradually decreased hydrophobicity is assumed to be represented by a three-dimensional Gaussian function:

\[
H_{tj} = \frac{1}{H_{tsum}} \exp \left( -\frac{(x_j - \bar{x})^2}{2\sigma_x^2} \right) \exp \left( -\frac{(y_j - \bar{y})^2}{2\sigma_y^2} \right) \\
\exp \left( -\frac{(z_j - \bar{z})^2}{2\sigma_z^2} \right)
\]

The value of \(H_{tsum}\) represents the sum of theoretical hydrophobicity of all the grid points. The value of the probability distribution (as the value of the Gauss function is usually interpreted) \(H_{tj}\) is assumed to represent the hydrophobicity density for the \(j\)-th grid point in the fuzzy-oil-drop. The hydrophobicity maximum is localized in the center of the ellipsoid and decreases in a distance-dependent manner according to the three-dimensional Gaussian function. The mean value at which the Gauss function reaches its maximum is localized at the (0,0,0) point in a coordinate system. The standard deviation represents the size of the drop (accoring the three-sigma rule) depending on the length of the polypeptide under consideration.

The protein molecule is localized with its geometrical center at the origin of the coordinate system. The longest distance between effective atoms (side chains represented by the geometrical centers of the atoms present there) determines the Z-axis. The Y-axis is oriented according to the longest distance between the projections of the effective atoms on the XY plane. The longest distance between the projections of the effective atoms along the X-axis determines the size of the drop.

For the orientation described above, the \(\sigma_x, \sigma_y, \sigma_z\) values can be calculated. The value of the longest distances versus the (0,0,0) point along each axis increased by 9 Å (cut-off distance for hydrophobic interaction) divided by 3 gives value of appropriate \(\sigma\).

A grid (its size depends on the molecule size) is created in three-dimensional space. The range of the Gauss function can be calculated for each grid point. The values of the Gauss function (representing the hydrophobicity distribution) are standardized to give a value of the sum of all values over all grid points equal to 1.0. The system of grid points calculated in this way is treated as the ideal fuzzy-oil-drop with a hydrophobicity distribution according to the Gauss function.
The Gaussian function makes it possible to calculate hydrophobicity in each point of space. The grid system can be created in step-wise form. In particular, the positions of effective atoms can also be treated as grid points, which collect the hydrophobic interaction of particular residue (j-th) with all others localized closer than the cutoff distance, and can further be treated as parameters describing a particular amino acid (j-th). This is why the distribution of hydrophobicity density can be presented in the form of a profile along the polypeptide.

Empirical oil-drop observed in proteins

The same grid points as described above are used to calculate the empirical hydrophobicity distribution represented by the empirical oil drop. The empirical hydrophobicity attributed to each grid point represents interaction with all effective atoms (below the cutoff distance). Generally, the grid point by itself represents zero hydrophobicity. If the grid point is localized in the place of effective atom of particular residue, its hydrophobicity is equal to the hydrophobicity of this residue. The observed hydrophobicity is calculated according to the simple sigmoid function previously proposed to quantitatively describe hydrophobic interactions [34]. The j-th point collects hydrophobicity H(o)j as follows:

\[
H_{oj} = \begin{cases} 
\frac{1}{H_{o,sun}} \sum_{i=1}^{N} (H_{i}^{t} + H_{i}^{c}) \left[ 1 - \frac{1}{2} \left( 7 \left( \frac{r_{ij}}{c} \right)^{2} - 9 \left( \frac{r_{ij}}{c} \right)^{4} + 5 \left( \frac{r_{ij}}{c} \right)^{6} - \left( \frac{r_{ij}}{c} \right)^{8} \right) \right], & \text{for } r_{ij} \leq c \\
0, & \text{for } r_{ij} > c 
\end{cases}
\]

where H(o)j represents the empirical hydrophobicity value characteristic for the j-th grid point, H(t)i represents the hydrophobicity characteristic of the i-th amino acid, rij is the distance between the j-th effective atom (generally j-th grid point) and i-th effective atom of the amino acid, and c expresses the cutoff distance, which has a fixed value of 9.0 Å following the original paper [34]. The value of H(o,sun) represents the sum of observed hydrophobicity of all the grid points.

Applying this function requires attribution of hydrophobicity parameters to each amino acid. Many scales for residue hydrophobicity are available. Some of them are based on analysis of known protein 3D structures [6, 14, 35–38], while others are derived from the physicochemical properties of amino acid side chains [39, 40]. Selection of an appropriate scale seems crucial, so a new statistics-based hydrophobicity scale for amino acids has been created.

The differences between idealized and empirical oil-drop

Since the theoretical H(t) and observed H(o) distributions of hydrophobicity are both standardized to 1.0, these two distributions can be compared. The differences between the theoretical and empirical distributions ΔH_i express the irregularity of hydrophobic core construction. For the i-th residue, ΔH_i is calculated as follows:

\[
\Delta H_i = H(t)_i - H(o)_i 
\]

where H(t)_i and H(o)_i are the theoretical and observed values of hydrophobicity for the geometric center of the i-th residue, respectively. The theoretical fuzzy-oil-drop and empirical oil drop were calculated for all proteins taken into consideration. A color scale was introduced to express the magnitude of the difference ΔH_i in a particular protein area, visualizing the localization of these discrepancies in the protein molecule. The one-dimensional profiles of ΔH_i were smoothened by averaging of the raw data using a five-residue running window frame. This simple method gives a legible curve without affecting the positions of dominant local extrema.

Comparative analysis

Method oriented on active site recognition SuMo [41] (http://www.sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome) was applied for comparative analysis. This method is based on the comparisons against binding sites from the PDB as criterion for active site recognition. The SuMo method is a two-step procedure. The representation of a protein structure by a set of chemical groups (unbound hydrogen bond donors or acceptors, accessible sides of aromatic rings and carboxylate, primary amide, etc.) is performed as a first step. The comparison of preformatted data is made in the second step. As a result, the list of similar active sites ordered according to decreasing score corresponding to the size of the matched sites is formed.

The results of SuMo and fuzzy-oil-drop-based were compared with the protein crystal structure. The distances below 12 Å between the ligand centre of gravity and Ca of sequential residues were used to define the amino acids responsible for ligand binding.
Fig. 1 One-dimensional profiles of $\Delta H$ per amino acid (left column) and three-dimensional distribution of $\Delta H$ on protein surface (right column) for lysozyme complexed with N-acetylglucosamine (a), chymotrypsin complexed with N-acetyl-L-phenylalanyl trifluoromethyl ketone (b), carboxypeptidase A2 inhibited by leech carboxypeptidase inhibitor (c), subtilisin DY inhibited by N-benzyloxycarbonyl-Ala-Pro-Phe-chloromethyl ketone (d), myoglobin complexed with heme (e) and ribonuclease complexed with guanosine-2'-monophosphate (f). The color scale expresses the magnitude of difference in a particular protein surface area. The dark blue (thick line) ligands are localized at their binding sites according to crystal structure. The 3-D structures were received using the VMD program with $\Delta H_i$ values put into the $\beta$-factor column and taken as criteria for the color scale.
Results

Comparing the expected and observed oil drops

The results of comparing the idealized and empirical hydrophobicity distributions are shown in two forms: a one-dimensional profile of $\Delta H_i$ (as dependent on the localization of the amino acid in the polypeptide chain), expressing the magnitude of differences between expected and observed hydrophobicity; and the three-dimensional distribution of $\Delta H_i$ (as it appeared on a particular part of the protein surface). Figure 1 shows the results describing six proteins.
selected for analysis of proteins complexed with small ligands. The left column represents the one-dimensional profiles of $\Delta H_i$. The color scale also visualizes these profiles. The same color scale adopted for the three-dimensional distribution of $\Delta H_i$ is shown in the right column. The dark blue (thick line) ligands, localized at their binding sites according to crystal structure, reveal very good agreement between the observed and predicted areas for potential interaction sites.

Fig. 2 One-dimensional profiles of $\Delta H$ per amino acid (left column) and three-dimensional distribution of $\Delta H$ on protein surface (right column) for cohesin-dockerin complex: results for cohesin (a), results for dockerin (b); complex between protein Ser/Thr phosphatase 1 and MYPT1: results for Ser/Thr phosphatase (c), results for MYPT1 (d); and LicT homodimer (e). The color scale expresses the magnitude of difference in a particular protein surface area.
Localization of biological function

Figure 2 also shows the localization of the area critical for the biological function of protein molecules. The proteins represent the case which can be called protein–protein complex construction. The profile of $\Delta H_i$ and its three-dimensional distribution as it appears in the protein complex are shown. Both partners’ binding sites can be easily recognized. The irregularity of the hydrophobicity distribution for the surface (hydrophobicity higher than the expected), and the cavity (hydrophobicity lower than expected) disturbing the regularity of the hydrophobicity distribution, seem to be good markers for mutual interaction.

Comparative analysis

The amino acids recognized as active site by SuMo and fuzzy-oil-drop model are given in Table 1. The residues identified on the basis of distance criterion are also given in Table 1.

The protein–protein complexes shown in the Fig. 2 have not been taken for this analysis due to SuMo limitation, which excludes the protein–protein complex creation.

The residues distinguished by fuzzy-oil-drop criterion are directly identified by maxima shown in Fig. 1. Two options are possible for the SuMo model: with and without ligand defined. To make the comparison complete, both forms were applied in this analysis. The score values are given in parenthesis. The residues identified on the basis of the crystal structure with distance (between C$\alpha$ position and ligand center of gravity). Figure 3 illustrates the comparison of the residues identified as belonging to active site.

The results of comparative analysis seem to represent rather high accordance particularly taking into account the character of the criteria used in these two methods. They seem to be quite differentiated: the protein molecule-oriented in fuzzy-oil-drop and ligand-oriented in SuMo.
Table 1 Numbers of residues recognized as belonging to active site according to: the close vicinity of residues versus the ligand center of gravity defined as distance below 10 Å and below 15 Å (numbers in parenthesis), according to SuMo in form of ligand defined and ligand not defined (the score values are given in parentheses) and according to fuzzy-oil-drop

<table>
<thead>
<tr>
<th>Protein</th>
<th>Crystal-based D&lt;10 Å (D&lt;15 Å)</th>
<th>SuMo Ligand defined</th>
<th>SuMo Ligand undefined</th>
<th>Fuzzy-oil-drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A6M</td>
<td>42, 43, 64–72, 89–93, 97–99, 103,</td>
<td>HEM: 92, 93, 97 (3.050)</td>
<td>HEM: 92, 93, 97 (3.050)</td>
<td>39–44, 68–71, 96–110,</td>
</tr>
<tr>
<td></td>
<td>104, 107 (25, 28–33, 39–46, 61–72,</td>
<td>HEM: 92, 93, 97 (3.050)</td>
<td>HEM: 92, 93, 97 (3.050)</td>
<td>134–141</td>
</tr>
</tbody>
</table>

Discussion and conclusion

Many proteins of unknown biological function, identified on the basis of genome analysis, await a unified automated method for determining their biological activity [42]. The next step is to develop methods able to predict a protein’s function from an examination of its structure. Some of the techniques used to identify functionally important residues from the sequence or structure are based on searching for homologues of proteins of known function [43–46]. However, homologues need not have related activity, particularly when the sequence identity is below 25%.
Geometry-based methods have shown that the location of active-site residues can be identified by searching for cavities in the protein structure [50] or by docking small molecules onto the structure [51]. Cavity localization in silico has been presented on the basis of the characteristics of the normal (perpendicular to the local surface) created for each portion of surface [52]. A complex analysis of protein interfaces and their characteristics versus highly divergent areas is presented in [53]. Several experimental studies have shown that mutations of residues involved in forming interfaces with other proteins or ligands can be replaced to produce more stable but inactive proteins [54–57]. On this basis, several effective algorithms have been developed [58, 59]. Finally, structural analysis coupled with measures of surface hydrophobicity has been used to identify sites on the surfaces of proteins involved in protein–protein interactions [60, 61].

The Critical Assessment of PRedicted Interaction (CAPRI) is oriented on blind prediction of protein complexes with ligands and protein–protein complex creation [62–66]. The results of competitions are presented on a web site (http://www.capri.ebi.ac.uk) and additionally published in Proteins Struc Func Gen (2003) 52 (the whole volume is devoted to this problem). The articles presented there show methods aimed at complex creation based mostly on detailed geometric surface analysis focused on the search for irregularities and cavities [67].

The method presented in this paper is especially dedicated for proteins of active site localized deep in the protein body. The amino acids of high negative values of ΔHi are understood as an area of higher than expected hydrophobicity. Those localized on the surface may suggest the area of potential protein interaction area. The amino acids of high positive values of ΔHi are interpreted as the area of lower than expected hydrophobicity. These fragments may be localized in close vicinity of the cavity ready to bind the ligand. The limitation of the described method is under consideration currently and will be published soon.

The results obtained by fuzzy-oil-drop model confronted with SuMo results and the data based on the crystal structure seem to be satisfactory. The best agreement is achieved for ribonuclease (1RGE) probably due to the high quality of crystallographic model and the diffraction data (measured by R-value) (R-value 0.109, resolution 1.15 Å). The high score given by SuMo appeared for the highest intermediate (I1). The number of intermediates (Ii) is unknown and presumably dependents on a particular protein or group of proteins. The final native structure (N) appears probably as the last one in a sequence of conformational modifications.

The conclusion given above may influence the construction of models for protein folding process simulation. The folding process recognized experimentally as multi-step process may be shown in a simplified as follows:

\[ U \rightarrow I_1 \rightarrow \ldots \rightarrow I_i \rightarrow \ldots \rightarrow N. \]

The unfolded structural form (U) is transformed into the first intermediate (I1). The number of intermediates (Ii) is unknown and presumably dependents on a particular protein or group of proteins. The final native structure (N) appears probably as the last one in a sequence of conformational modifications.

The model developed (part of which is presented in this paper) assumes the two-step process with following intermediates:

\[ U \rightarrow ES \rightarrow LS \rightarrow N \]

The unfolded state (U) representing the starting event adopts the early-stage conformation (ES), which is determined by backbone conformation presented formerly [81–83] and applied to folding simulation of few proteins [68, 83–85]. The sequence-to-structure (ES) relation generalized.
in form of contingency table allowed application to any protein under consideration [69, 86, 87].

The ES form changed to late-stage (LS), which is mostly driven by hydrophobic interaction between side chains is supposed to approach the native structure (N) of the protein.

The LS conformation of protein can be created in silico according to the fuzzy-oil-drop model [72, 74, 88]. The results presented in this paper (as well as other results [73, 75, 89, 90]) suggest that the active presence of ligand during the protein folding process influencing mutually conformation of both molecules may ensure creation of highly specific ligand binding cavity [73, 75].

In consequence, the simulation of protein folding process can be presented as follows:

\[ U \rightarrow ES \rightarrow LS \rightarrow N \]

Initial results [72–75, 88–90] encourage large-scale computational experiments, suitable for the grid environment. Further simulations are planned as a part of the “Never Born Proteins” project in the application layer of the EUChinaGRID initiative.

The procedure for folding simulation according to the presented model requires the knowledge of ligand molecule properties (e.g., their electronic structures, partial charges, van der Waals parameters). Presently, such detailed chemical characteristics of ligand (substrate) molecules is not available in a consistent form. Therefore, one of our goals is to create the library delivering the parameters describing the molecules found as protein ligand in the Protein Data Bank (PDB) [91]. The ligand properties are calculated using quantum chemical methods. The optimal geometry of ground state and the atomic partial charges of isolated ligand molecules as well as their energy are calculated using the Amsterdam Density Functional (ADF) [92] in the TZV extended basis set (Slater type functions) with BP86 exchange correlation functional. The atomic charges are calculated according to Mulliken population analysis [93] and Hirshfeld analysis [94]. We plan, also, to enclose similar properties computed by Gaussian package [95] for comparison (e.g., partial charges implying from the fitting of electrostatic potential [96]). The service is progressively available on the http://www.bioinformatics.cm-uj.krakow.pl/ligand website.

Availability

The active site can be easily predicted for any protein structure according to the presented model, with the free prediction server available at http://www.bioinformatics.cm-uj.krakow.pl/activesite.

Acknowledgements This research was supported by the Polish State Committee for Scientific Research (KBN), grant 3 T11F 003 28, and Collegium Medicum grants 501/P/133/L and WL/222/P/L. The work has been supported by the European Commission EuChinaGRID project (contract number 026634). Chemical quantum calculations for ligand database were performed at the Research Centre in Juelich (project no. 2249) – ADF and at the Academic Computer Centre Cyfronet AGH - Gaussian. Many thanks to Olga Stepien for the implementation of the web interface.

References
