Kochańczyk BMC Structural Biology 2011, 11:34
http://www.biomedcentral.com/1472-6807/11/34

BMC Structural Biology

RESEARCH ARTICLE Open Access

Prediction of functionally important residues in globular proteins from unusual central distances of amino acids

Marek Kochańczyk[1,2]

Abstract

Background: Well-performing automated protein function recognition approaches usually comprise several complementary techniques. Beside constructing better consensus, their predictive power can be improved by either adding or refining independent modules that explore orthogonal features of proteins. In this work, we demonstrated how the exploration of global atomic distributions can be used to indicate functionally important residues.

Results: Using a set of carefully selected globular proteins, we parametrized continuous probability density functions describing preferred central distances of individual protein atoms. Relative preferred burials were estimated using mixture models of radial density functions dependent on the amino acid composition of a protein under consideration. The unexpectedness of extraordinary locations of atoms was evaluated in the information-theoretic manner and used directly for the identification of key amino acids. In the validation study, we tested capabilities of a tool built upon our approach, called SurpResi, by searching for binding sites interacting with ligands. The tool indicated multiple candidate sites achieving success rates comparable to several geometric methods. We also showed that the unexpectedness is a property of regions involved in protein-protein interactions, and thus can be used for the ranking of protein docking predictions. The computational approach implemented in this work is freely available via a Web interface at http://www.bioinformatics.org/surpresi.

Conclusions: Probabilistic analysis of atomic central distances in globular proteins is capable of capturing distinct orientational preferences of amino acids as resulting from different sizes, charges and hydrophobic characters of their side chains. When idealized spatial preferences can be inferred from the sole amino acid composition of a protein, residues located in hydrophobically unfavorable environments can be easily detected. Such residues turn out to be often directly involved in binding ligands or interfacing with other proteins.

Background

The task of assigning a function to each new protein structure resulting from high-throughput structural genomics experiments requires reliable computational annotation methods. Identified functionally important amino acids can provide preliminary clues on the co-evolution and molecular workings of proteins. Such information is crucial for the site-directed mutational engineering and *de novo* protein design. The integration of knowledge of the locations of binding sites with ligand screening or docking protocols improves initial stages of the rational drug design [1]. Also, when putative residues responsible for the complex formation are identified, protein-protein interaction interfaces can be characterized *in silico* [2].

Currently, due to the availability of 3D data, the exploration of properties embedded in the structure of proteins prevails over the traditional motif recognition and sequence comparison (that may turn out to be surprisingly ambiguous [3]). For close homologs, the knowledge-based approaches transfer functional annotations from proteins with already known structure and function [4-8]. Their average effectiveness is inherently limited by the availability of solved and annotated

Correspondence: marek.kochanczyk@uj.edu.pl
[1]Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, ul. Reymonta 4, 30-059 Krakow, Poland
Full list of author information is available at the end of the article

structures, so more generic methods are still desirable. Numerous pure geometry-based methods search locally for clefts and pockets in the molecular surface by employing computational geometry algorithms [9-16]. The spatial neighborhood of residues is used to characterize local environments in methods that take into account additional factors such as the flexibility of residues [17], electrostatic potential [18,19] or overall interaction energy [20], excess or deficiency of the hydrophobicity [21], hydrophobic potential around a protein [22] or a multitude of other, predominantly physicochemical, residue properties [23-27].

Interestingly, indications based on diverse descriptions are usually not correlated [28]; nor can they be used for the prediction of both protein-ligand and protein-protein interaction sites [29]. As a consequence, well-performing present-day approaches use combinations of complementary characteristics, for example the electrostatics and geometric properties [30] or the geometry and conservation [31-33]. Metaservers offer combinations of several independent fully-fledged methods in order to compensate for the shortcomings of some methods with capabilities of others [34,35]. As the compositions of distinct binding site prediction methods achieve better success rates than constituent techniques applied solo, it is still valuable not only to provide fine-tuned variations of heterogeneous approaches, but also to search for assorted methods that could complement existing ones by the exploration of specific orthogonal features.

Contrary to the majority of approaches that characterize fragments of proteins locally and with a considerable degree of detail, Brylinski et al. [21,36] showed that the rough analysis of the global spatial distribution of amino acids with respect to their hydrophobicity is capable of localizing ligation sites. They did not follow usual hydrophobicity quantifications such as the average solvent-accessible surface area or number of contacts [37], but rather measured the discrepancy between idealized and observed hydrophobicity within the fuzzy oil drop model [38], where the trivariate Gaussian distribution is used to express the idealized protein hydrophobicity (maximum value in the protein core, smoothly approaching 0 about and beyond the perimeter). It turned out that amino acids of high discrepancy (unexpectedly high hydrophobicity in relation to their peripheral position) often occur in function-related areas of proteins.

This observation is fundamental to the current work, where we devised and validated a method for the identification of function-related residues based on the probabilistic description of atomic burials originating from the conceptual framework of Gomes et al. [39]. We collected necessary statistics from a selection of globular proteins and, as opposed to the original application of the framework, we used a radial probability density function to describe preferred central distances of individual atoms of types defined within amino acids. In this view, proteins are treated as mixtures of amino acids where restraints resulting from their covalent connectivity are ignored (except for cysteines). Any deviations from the spherical shape of the macromolecule, intrinsic rigidness imposed by the presence of secondary structures and local interactions are neglected: proteins are treated as compact solid-like bodies of atoms, where the isotropic hydrophobic segregation and packing are considered to be the dominant driving forces conferring spatial organization of residues [40-42].

The classic analysis of just several protein structures suggested that the sole orientational preferences of side chains can be a criterion for the hydrophobic or hydrophilic character [43]. Therefore, although a multitude of hydrophobicity scales or burial indices are available for (whole) amino acids and many knowledge-based pair-potentials are constructed for (united) residue side chains [44], we decided to act on the per-atom rather than per-residue basis in order to account for (radial) orientational preferences of residues. The actual amino acid composition of a protein influences its native structure topology [45,46], folding type [47,48] and interactions [49]. In our statistical model, for a protein with a known amino acid abundance we assume that the relative probabilities are directly proportional to the stoichiometry. In our approach to the function prediction, every heavy atom in every amino acid of the protein considered has the measure of its *unexpectedness* estimated with respect to all possible atom types in a given point of space. The measure depends solely on the distance from the geometric center of the polymer. Typically, residues that place their atoms in the least probable central distances appear to contribute to the creation of ligand binding sites (including active sites of enzymes) or protein-protein binding interfaces.

Methods

Extraction of a non-redundant set of globular proteins

We examined a total of 172 265 protein chains as deposited in RCSB PDB [50] in January 2011 and excluded structures of high asymmetry or in other aspects irregular. Two geometric descriptors were used discriminatively: asphericity, calculated as the normalized sum of squared differences of the eigenvalues of the gyration tensor (according to [51]), was required to be smaller than 0.1 and compactness to be at least 0.5; the latter value was calculated as the ratio of the solvent accessible surface area of the (ideal) sphere of the volume of a considered protein to its actual solvent accessible surface area (this is a more intuitive inverse of the fraction introduced by Galzitskaya et al. [52]). Chains of sequence lengths smaller than 100 amino acids were excluded due to strong geometric constraints. Proteins that fulfill all the aforementioned conditions are denoted as globular in this paper.

Furthermore, it was required that every solved structure should contain no discontinuities, be determined with an experimental method to a resolution better than 2 Å, contain only a single domain (according to both SCOP [53] and CATH [54] classifications) and must not create multi-chain complexes, even transiently (determined on the basis of biological units assemblies available from PDB). A total of 2953 proteins were extracted for further considerations (1.71% of the whole PDB).

In the last step, in order to reduce sequence redundancy, precomputed clustering results available from the PDB, generated by the Cd-hit program [55] that grouped sequences of at least 90% of sequence identity in clusters, were used to select a single protein per every cluster. Finally, the learning data set comprised 775 high-resolution single-domain globular chains (26.2% of previously selected chains). The full list of PDB ids is available in Additional file 1 Table S1.

Compactness and asphericity of proteins in the set turned out to be only weakly interdependent (correlation coefficient, CC, -0.14). Longer chains were characterized by lower compactness (CC = -0.45) but not necessarily higher asphericity (CC = -0.06). Distributions and dependencies of geometric descriptors are presented in the Additional file 2 Figure S1.

Probabilistic description of atomic burials

Geometric centers and radii of gyration were calculated for every chain in the learning set. Distances to the geometric center of a chain of every heavy atom, r, were divided by the radius of gyration of the whole chain, r_g, enabling a uniform view of globular proteins of various sizes [43]. Histograms of such normalized distances, $R = r/r_g$, were collected for every amino acid-dependent atom type denoted by τ. Three types of cysteines were considered separately: generic Cys (irrespective of the presence or absence of SS bonding), Cys creating (intrachain) disulfide bridges (denoted CSS, nearly 40% of all Cys) and Cys reduced and not involved in SS bridging (C_{SH}). A total of 170 histograms for different τ were obtained.

A continuous "mass" function derived by Gomes et al. [39] to describe burials of whole residues was considered for fitting. The original function expresses the quadratic increase of the volume when moving away from the core of a protein and sigmoidal decrease (Fermi function) of the atomic density in the rim as dependent on the normalized radius, R:

$$p_\alpha(R;\tau) = \frac{A_\tau R^2}{1+\exp(\beta_\tau(R^{\alpha_\tau}-\mu_\tau))}. \tag{1}$$

After applying the direct least-squares method for fitting individual histograms, obtained fits yielded unsatisfactory sums of the squared residuals (SSR) for atoms in hydrophilic residues, where the expression overestimated their propensity to occur in the protein core. To account for this observation, the assumption of the strictly quadratic increase was abandoned and an additional tunable parameter, γ_τ, was introduced while α_τ was set to 1 (see Additional file 3 Figure S2). The following form was finally used:

$$p(R;\tau) = \frac{A_\tau R^{\gamma_\tau}}{1+\exp(\beta_\tau(R-\mu_\tau))} \tag{2}$$

for fitting. Parameter A_τ provides normalization, μ_τ principally determines location, β_τ influences the width of the distribution and γ_τ controls convexity of the left ridge. The goodness-of-fit of distributions of the latter form was better for 124 of 170 fits (in terms of SSR) in comparison to the original distribution function with variable α (Equation 1) and for 130 of 170 fits (F-test with p-value < 0.000001) in comparison to the original distribution function with $\alpha = 1$.

Expected atomic burials in proteins

Densities of atoms are characterized globally in the environment of the protein itself in the common and reduced coordinate space. Thus, assuming the lack of void spaces inside, in a given point in space, located in the normalized distance R from the geometric center of the protein, one can estimate the expected chance of occurrence of an atom τ by relating its probability, $p(R;\tau)$, to probabilities of occurrences of all atoms, $\Sigma_{\tau\in T}\, p(R;\tau)$, where T is the complete set of 170 atomic types. As we consider concrete protein species, probabilities depend effectively on the number of atoms τ (equal to the number of amino acids of a concrete type) present in the whole protein, $n(\tau)$. Only their relative fractions are important so we can use them directly for weighting in the expression similar to the posterior distribution of component membership in mixture models. The equation

$$\bar{p}(R;\tau) = \frac{n(\tau)p(R;\tau)}{\sum_{\tau'\in T} n(\tau')p(R;\tau')} \tag{3}$$

is used for the estimation of expected atomic central distances in proteins with known amino acid composition. The variability of preferred atoms in a given point in space is measured in bits as the entropy of expected burials:

$$S(R) = -\sum_{\tau\in T}\bar{p}(R;\tau)\log_2\bar{p}(R;\tau). \tag{4}$$

Prediction of functionally important residues

In search of residues employed directly in performing the function, we follow the crucial observation by Brylinski et al. [56] that irregularities in the global distribution of hydrophobicity often indicate function-related areas. We follow this principle in our probabilistic approach by searching for atoms of the relatively least probable central distances, $\bar{p}(R;\tau)$. Residues with such atoms are usually the hydrophobic amino acids exposed to the solvent or hydrophilic amino acids located close to the protein core. The unexpectedness of a central distance can be converted into a simple free energy-like term by the following equation:

$$\text{Unexpectedness}(R;\tau) = -\log_2 \bar{p}(R;\tau), \quad (5)$$

which gives estimates in bits.

Prediction of ligand binding sites

As for compact structures it holds that r_g is roughly proportional to (sequence length)$^{1/3}$ [57] and as in the task of binding sites recognition one is interested primarily in non-buried residues on the surface, the area of which is proportional to r_g^2, as a rule of thumb, $\left\lfloor \frac{1}{4} \cdot (\text{sequence length})^{2/3} \right\rfloor$ residues containing the most unexpected atoms are initially selected. (However, assuming the general spatial character of the statistical model, no additional factors such as estimates of solvent accessibility are taken into account.) Selected residues are weighted proportionally to the maximum value of unexpectedness among values assigned to constituent atoms and then clustered hierarchically using the pairwise average-linkage method. In search for ligand binding sites, the hierarchy of residues is partitioned into clusters separated by more than 7 Å (average Euclidean distance) that indicate (possibly multiple) putative sites. Positions of cluster centroids are computed in a weighted manner and located closer to the most unexpected atoms. Putative sites are ranked according to the proximity of their predicted centroids to the geometric center of the whole protein.

Prediction of protein-protein interfaces

Contrary to the development of the complete algorithm for the prediction of binding sites of (small) ligands, we do not attempt to create a new protein-protein docking method but rather to provide a simple unexpectedness-based scoring function for the ranking of docking predictions. Heavy atoms of one protein located within a distance of 10 Å from the other have their unexpectedness calculated and a maximum value of unexpectedness is found in this way for both macromolecules of a docked assembly. A docking prediction is then scored using the average of the highest values of unexpectedness in two interfaces.

Evaluation of predictions

The evaluation of the method based on the introduced characteristics was performed separately for the task of predicting binding sites of small ligands and for the prediction of regions creating interfaces to other proteins. In both cases, if a test data set allowed, predictions were made for unbound structures; after the assignment, the *apo* form was superimposed onto the *holo* form so that intermolecular distances were measured between the unbound structure and ligand/another macromolecule as located in the structure of the complex.

For the prediction of ligand binding sites, a set of 48 pairs of unbound/bound structures and a set of 210 bound structures, which were already employed for the benchmarking of other methods (LigSite^{csc} [32] and IBIS [8]), were used for the comparison with already measured success rates of the state of the art geometry-based methods: SURFNET [9], PASS [10] and LigSite [12]. The former set, further referred to as the LB_{48} test set, includes 38 enzymes that cover 39 diverse enzymatic activities according to the EC annotations from the Catalytic Sites Atlas version 2.2.12 [58] and 10 proteins that bind compounds in their non-active sites. The latter set, referred to as the LB_{210} test set, enabled large-scale benchmarking.

In order to juxtapose the results of our approach and similar fuzzy oil drop-based method (FOD), which assign prediction scores to clusters of atoms, with pocket identification methods, which indicate geometric centers of pockets located over the molecular surface, we used MSMS [59] and projected coordinates of centroids of putative binding sites onto the solvent-excluded molecular surface. Then, in order to apply the cut-off value of 4 Å used in pocket prediction benchmarks, we displaced surface-projected coordinates by 1 Å in the direction of the vector normal to the surface and 1 Å outwards from the geometric center of the protein. As the points do not always lie the space in the pocket, additionally we used the cut-off of 6 Å. We examined whether any atom of the ligand is located within the cut-off distance and reported success rates for the best ranked (Top 1) and 3 highest ranked (Top 3) candidate sites.

In order to show, preliminarily, that the unexpectedness is a property of protein-protein interfaces, we used the latest and most extensive docking benchmark (version 4.0) [60], further referred to as the PPI_{176} test set. Residues of two macromolecules were considered as interfacing if they were separated by at most 4 Å. In the case of protein-protein binding interfaces, unexpected residues are usually isolated, so we did not cluster them, but rather reported the average unexpectedness in binding/non-binding protein regions.

Eventually, the capability of appropriate ranking of protein-protein docking predictions was compared to that of one of the best performing docking algorithms, ZDock [61], optionally amended with ZRank [62], and two other methods, recent ASP-Dock [63] and older FTDock [64]. The methods have their success rates already measured over the complete protein docking benchmark version 3.0 [65], so this set (referred to as the PPI_{124} test set) was used to estimate the capacity of our approach. The unexpectedness-based score assessed 54,000 docking poses of a decoy generated by ZDock 3.0 operating at the rotational scanning interval of 6°. A successful prediction was defined as a docking solution of ligand C^{α} RMSD < 10 Å.

Comparison with other characteristics

A direct evaluation of the current method was performed in parallel with the fuzzy oil drop (FOD) method [21] using the LB_{48} test set. The same clustering and ranking methods were used for residues with the highest unexpectedness and for residues of the highest observed vs. theoretical hydrophobicity discrepancy, $\Delta\tilde{H}$ (FOD). For the detailed comparison with other explorable characteristics, useful for the prediction of (small) ligand binding sites, the evolutionary conservation scores were assigned to residues according to the multiple-sequence alignment-based ConSurf-DB [66]; only residues of the highest conservation score (i.e. 9) are indicated in this paper. Independently, the clusters of ionisable residues with anomalous predicted titration behaviour, identified with the finite difference Poisson-Boltzmann-based technique, Thematics [25], were included in the comparison.

Results

Orientational preferences of amino acids

Parameters of probability distribution functions given by Equation 2, A_τ , μ_τ β_τ and γ_τ, were determined independently for every amino acid-dependent atom type, τ, allowing to capture the specific radial orientational propensities of amino acids. The full list of 170 sets of parameters for atomic distribution functions derived from the obtained learning set can be found in the Additional file 4 Table S2. Since the structure of side chains allows to single out the atom most distant from the C^{α} atom, it is possible to capture and demonstrate preferred orientations using a less redundant description. We decided to evaluate unexpectedness of every atom uniformly motivated by the fact that among 83 distributions of all side chain heavy atom types as many as 58 were statistically significantly different than distributions of relevant C^{α} atoms (Kolmogorov-Smirnov tests with p-value < 0.000001; see Additional file 4 Table S2 for details).

Resulting probability density functions have nonzero skewness, so in order to portray synthetically the orientational preferences, we use both differences between mean values and between maxima of distributions of C^{α} and distal atoms (Figure 1). The arrows can be interpreted as expressing global hydrophobic moments of (amphiphilic) residues defined in the environment of the protein itself (analogous to [67]). In this view, the two amino acids of the most prominent opposite orientational preferences are Lys and Phe (Figure 2).

Although side chains determine the hydrophobic/hydrophilic character of amino acids, they influence considerably probabilities of spatial occurrence of (chemically equivalent across amino acid types) C^{α} atoms. In the synthetic picture of atomic densities (Figure 1 and Additional file 5 Figure S3), hydrophobic propensities of amino acids in the body of a protein are modulated by their sizes: broad distributions of Gly and Ala atoms are shifted from those of other hydrophobic types; distributions of large amino acids, such as Trp or Arg, are less dispersed around their maxima; the broad distribution of His can be explained by diverse possible protonation states and the ambivalent distribution of Tyr - by mixed aromatic/polar character of its side chain.

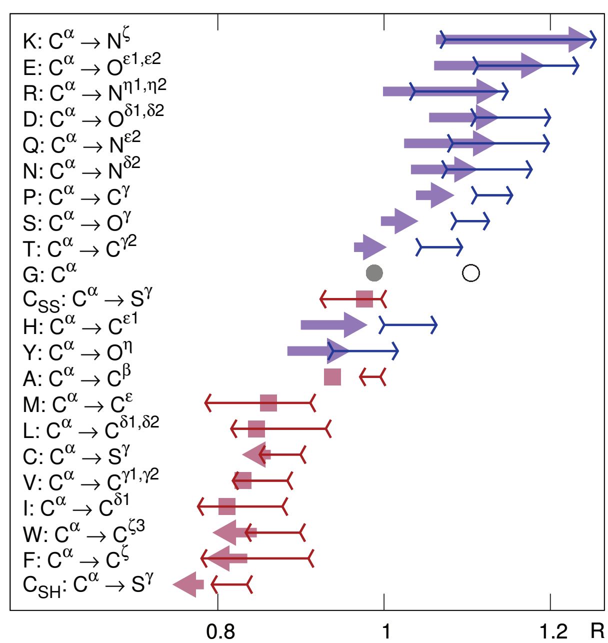


Figure 1 Orientational preferences of amino acids in globular proteins. Locations of mean and maximum values of probability density functions for C^{α} and most distant side chain atoms for all amino acids. Thick arrows connect means; thin arrows span between maxima of distributions. All arrows point towards the most distal atom in the side chain (except for Gly) according to the labels on the left. The arrows that would be shorter than their heads are replaced by squares.

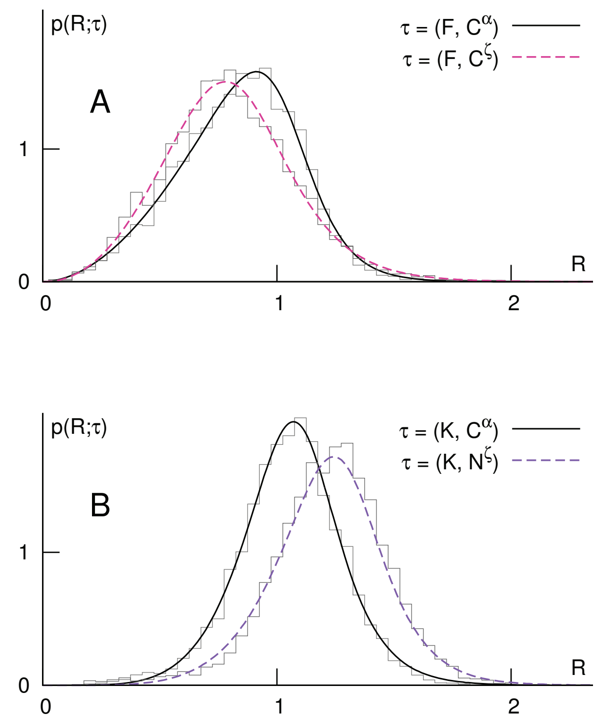


Figure 2 Probability densities of C^{α} and of the most distal side chain atom in Phe and Lys. The two amino acids exhibit the most prominent (the largest separation of the means) and opposite (centripetal, **A**, vs. centrifugal, **B**) orientational preferences.

The analysis of the intriguing case of Cys reveals that, although their orientation does not depend on the possible disulfide bonding, the non-bridging cysteines prevail as the most buried residues, while those constituting cystines occur more often on the protein surface (Figure 1; Additional file 6 Figure S4). Cysteines are relatively frequently found in active sites [68]; supposedly, the evolution may easily redefine the function of a protein by tailoring the state of cysteines and adjusting their positions [69].

Distribution of unexpectedness

The mean central reduced distances of distal site chain atoms are in agreement with known hydrophobicity scales, especially those empirical ones based on the surface accessibility. Several theoretical and one experimental scale, along with similarities expressed in terms of the correlation coefficient, are listed in Table 1.

The statistical model applied to globular proteins from the learning set reveals a critical value of about $0.93 \cdot r_g$, where the average entropy, calculated according to the Equation 4 and interpreted as the lack of preference for particular atomic types, has the highest value (Figure 3). The value marks clearly the hydrophobic-hydrophilic transition on the protein surface, usually covered by a patchwork of hydrophobic and hydrophilic areas [70,71]. Although it was observed in larger proteins that the degree of hydrophobicity is constant for $R < 0.7$ [72], according to the model the protein interior is not a volume of uniform preferences, but rather it visibly exhibits a gradually increasing preference for some apolar atomic types (decreasing entropy) when moving towards the centroid.

Types of the most unexpected amino acids (i.e. amino acids comprising most unexpected atoms) were determined in the LB_{48} test set and in the PPI_{176} test set separately (Figure 4). In the former set, the additional requirement of $R < 0.93$ and in the latter the requirement of $R > 0.93$ were imposed, because several proteins in the LB_{48} test set create complexes with other proteins and proteins in the PPI_{176} test set contain ligand binding pockets. According to the model, the most unexpected residues lying within the radius of gyration are those charged or ionizable, such as Glu, Asp, Lys and Arg, which are known to play essential functional roles in the enzymatic active sites. Amino acids with branching aliphatic side chains, Leu, Val and Ile, are properly assessed as being rarely exposed to the solvent. Unfortunately, broad distributions of central distances of His and Tyr cause them to be hardly ever indicated as unexpected. Also, due to the specific structural roles of Pro and Cys, such residues tend to be rated as unexpected despite the possible lack of any direct relation to the function.

Prediction of ligand binding sites

Clusters of unexpected residues turn out to be located on the surface of proteins, very often inside clefts and

Table 1 Correlations of mean values of distal side chain atom distributions to other characteristics

CC	Description of the characteristics	Reference
-0.984	Mean fractional area loss upon folding	[88]
-0.974	Solvent accessibility based on self-information [16% accessibility]	[89]
-0.971	Information value for accessibility [average fraction 35%]	[90]
+0.961	Normalized eigenvector of the Sweet & Eisenberg scale	[91]
-0.951	Mean combined polarity calculated from distributions of residues in proteins	[92]
+0.897	Hydrophobicity coefficient in RP-HPLC [C4 with 0.1%TFA/MeCN/H_2O]	[93]

Similarities of 5 theoretical (top) and 1 experimental (bottom) single-value amino acid characteristics are expressed in terms of the correlation coefficient, CC. (For Cys, the distribution of reduced S^{γ} was used.)

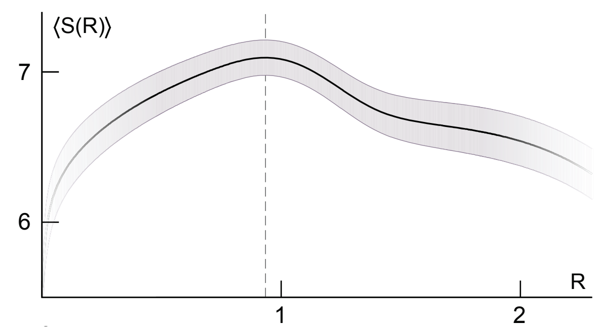


Figure 3 Entropy of expected reduced central distances in globular proteins. The entropy, *S*(*R*), of $\bar{p}(R;\tau)$ was averaged over all proteins from the learning set and is expressed in bits (black line; gray band - standard deviation of the mean entropy; dashed line - location of the maximum). "Twilight zones" mark regions where the entropy was calculated using tails of distributions.

pockets, where ligand compounds are bound. Geometric centroids of such clusters designate candidate ligand binding sites with the success rate similar to that of the fuzzy oil drop-based method in the LB_{48} test set and only slightly worse in the LB_{210} test set (see Table 2). For the cut-off value of 6 Å of the distance to a ligand, considered as enabling the comparison, the performance of both global hydrophobicity distribution-based strategies is similar or even marginally better than that of three state of the art methods, PASS, LIGSITE and SURFNET, which distinguish clefts or cavities based solely on the local geometry (Table 2).

The relations to other characteristics frequently exploited for the localization of binding sites, viz., conservation and electrostatics, were examined for residues in properly indicated Top 3 clusters (Table 3). There are no clusters with active site residues displaying neither

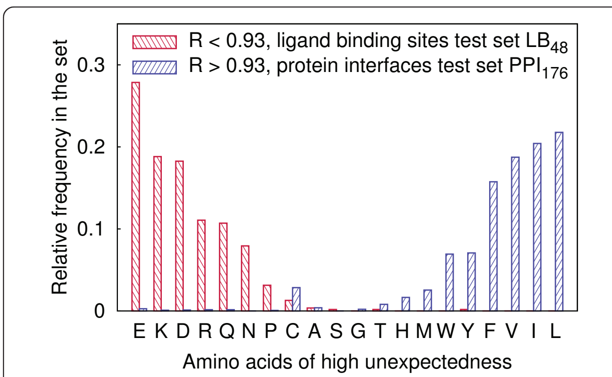


Figure 4 Relative frequencies of amino acids characterized by high unexpectedness. Residues lying within $0.93 \cdot r_g$ in proteins from a set used for the ligand binding site prediction and residues of central distances greater than $0.93 \cdot r_g$ from a set used for the protein-protein interface prediction are presented separately.

Table 2 Benchmarks of several ligand binding site prediction methods

	LB_{48} test set		LB_{210} test set	
Method	**Top 1**	**Top 3**	**Top 1**	**Top 3**
PASS	60*	71*	54*	79*
LIGSITE	58*	75*	65*	85*
SURFNET	52*	75*	42*	56*
FOD	56 (71)	60 (81)	55 (68)	72 (83)
Unexpectedness	48 (69)	63 (83)	53 (65)	67 (80)

The comparison of ligand binding site prediction success rates of the current approach (Unexpectedness), a global hydrophobicity-based method (FOD) and several non-hybrid pocket-searching state of the art methods for 48 unbound molecules from the LB_{48} test set. The cut-off distances are 4 Å and 6 Å (success rates for the latter value are in parentheses). Results marked with stars were reported in [32].

Table 3 Residues in correctly predicted 3 top-ranked clusters

Structure	Function	Cluster
1ahc A	(RNA) glycosidase	**R**, **E**, E, Q
1bbs A	proteinase	**D̈**, **D̈**
1bya A	O-glycosidase	**Ë**, R, Ë, P
1cge A	metalloproteinase	**Ë**
1djb A	hydrolase (*β*-lactamase)	**K**, **E**
1hsi A	protease (HIV-2 retropepsin)	I (flaps)
1hxf H	(serine) protease	**D̈**
1ifb A	*fatty acid binding*	R̈, E
1ime A	(inositol) phosphatase	D̈, D̈, D̈
1krn A	hydrolase (fibrinolysin)	K̈
1l3f E	proteolysin	**E**
1nna A	O-glycosidase	K̈, **Ë**, R, Ë, Q
1npc A	(metallo)protease	**E**, E
1pdy A	enolase	**K**, R, Q
1psn A	(acid) proteinase	**D̈**, **D̈**
1pts A	*azobenzoic acid binding*	D
1qif A	(acetylcholin)esterase	Ë
1stn A	(phosphodi)esterase	**R**, D̈
1ypi A	(triosephosphate) isomerase	**K**, E
2cba A	lyase (anhydrase)	Ë, **Ë**
2ctb A	hydrolase (carboxypeptidase)	**E**
2fbp B	(fructose bis)phosphatase	K, Ë, D, D̈, Ë
2sil A	hydrolase (neuraminidase)	Ë, Q, Q, R̈, R̈
3app A	(acid) proteinase	**D̈**
3p2p A	(carboxyl)esterase	R, **D**
3ptn A	hydrolase (tripsin)	**D**
3tms A	(methyl)transferase	**Ë**, **N**, Q
5dfr A	(folic acid) reductase	**D**
8adh A	dehydrogenase	E, D̈
8rat A	hydrolase (ribonuclease)	**K̈**, Q

Residues are sorted in rows according to decreasing unexpectedness. Residues of the highest evolutionary conservation scores according to the ConSurf-DB [66] are underlined; residues indicated as functional by Thematics [25] have overbars; bold residues are annotated as catalytic in the Catalytic Sites Atlas (CSA) [58]. (Two chains of non-enzymatic functions are unannotated in the CSA.)

conservation nor the indicative anomalous ionisable behavior - in fact, in most cases there is a significant overlap between the unexpectedness and two other attributes; in remaining cases the three features may be seen as complementing one another (especially for residues that are nonionizable or bind with low specificity).

Among the proteins annotated with EC numbers in the LB_{48} test set, 35 out of 38 enzymes have their active sites recognized in Top 3 clusters (31/38 in Top 1). Notwithstanding, out of 10 proteins that exhibit no enzymatic activity and bind ligands in their non-active sites, binding sites are properly recognized in only 5 cases, mainly because of their eccentric locations (see Additional file 7 Table S3 for details).

The predictive power of our approach decreases moderately for more aspherical proteins. The quality of cluster rankings seems to be independent of the asphericity (Figure 5).

Ranking of protein-protein docking results

The unexpectedness was employed to characterize the protein-protein interfaces in the PPI_{176} test set, where the majority of structures have the asphericity higher than 0.1. Despite this difficulty, the median unexpectedness of interacting residues turns out to be clearly higher than the median unexpectedness of all surfaces residues (Figure 6). When a subset of more globular proteins is examined, the difference is even more salient (not shown).

Scoring of interfaces based on the unexpectedness yields consistently better results than an analogous FOD-based scoring for 100 top-ranked solutions (Figure 7). For 10 top-ranked docking solutions success rates of our approach are nearly comparable to that of the ZRank, indicating that our score can properly account for desolvation and electrostatics-related properties used (in addition to van der Waals interactions) by ZRank.

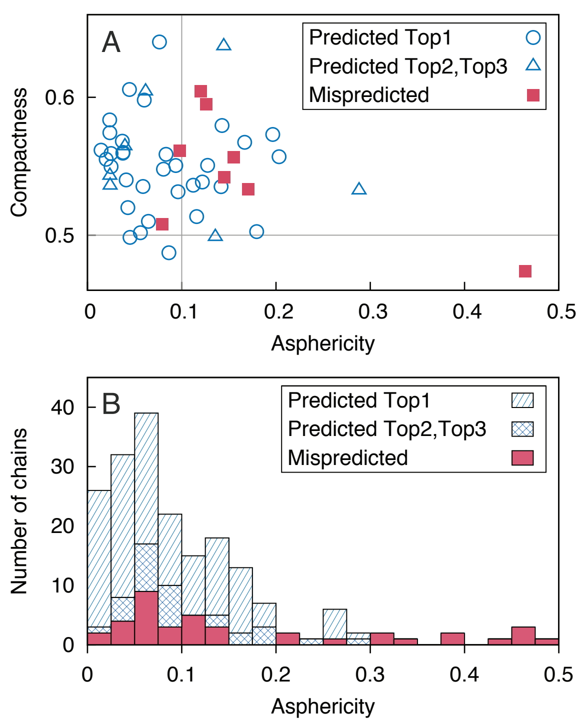


Figure 5 Dependence of the prediction success on geometric characteristics. Predictions for apoproteins from the LB_{48} test set (**A**) and holoproteins from the LB_{210} test set (**B**) are shown (cut-off 6 Å). Two perpendicular thin gray lines correspond to the geometric requirements imposed on proteins in the learning set.

Comparison to the fuzzy oil drop model

Ranking clusters according to the most unexpected atoms turned out to be less specific than the ordering based on the FOD-based discrepancy between theoretical and empirical hydrophobicity, $\Delta\tilde{H}$. Searching for the reason of disadvantageous cluster rankings we found that the FOD method not only quantifies the hydrophobicity discrepancy, but primarily indicates residues in the proximity to the molecular centroid (Figure 8). Visibly, the fuzzy oil drop model inadequately overestimates the hydrophobicity in protein cores. The satisfactory predictive capability and advantageous ranking of the FOD-based method can be explained by the observation that the distance to the centroid can be used autonomously for the detection of active sites and enzyme-ligand interfaces [73]. In our probabilistic approach, unexpectedness of atoms is virtually independent of their central distances.

Availability

We developed a web server SurpResi for the prediction of functionally important sites based on the unusual

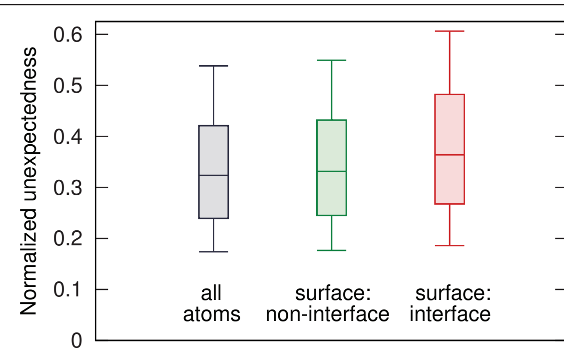


Figure 6 Unexpectedness of atoms in protein interfaces. Normalized values of unexpectedness for all atoms and atoms belonging to residues exposed to the solvent (residue solvent-accessible area >10 Å^2), subdivided into these creating and not creating protein-protein interfaces. Whiskers represent the 9th and the 91st percentile.

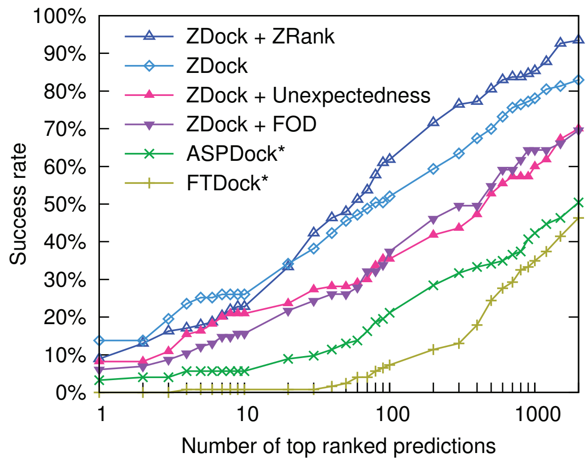


Figure 7 Effectiveness of rankings of docking solutions. Decoys generated and ranked by ZDock were reranked using ZRank, $\Delta\tilde{H}$ (FOD) and the unexpectedness. Success rates of two independent docking approaches, FTDock and ASPDock, marked with stars, are displayed as reported in [63].

central distances of atoms. The input of SurpResi server is a Protein Data Bank (PDB) file or user file in the PDB format. The output is a downloadable PDB file where the column of beta factors is replaced by the unexpectedness and the occupancy is replaced by the same value normalized to the range [0,1] over all protein atoms. In the header section, the file contains detailed information about clustering and ranking of clusters. The web server and source code are freely available at http://www.bioinformatics.org/surpresi.

Discussion

The presented approach quantifies polar and directional propensities of amino acids using the partition in the knowledge-based continuous gradient of hydrophobicity generated by the protein itself. It yields a middle level of

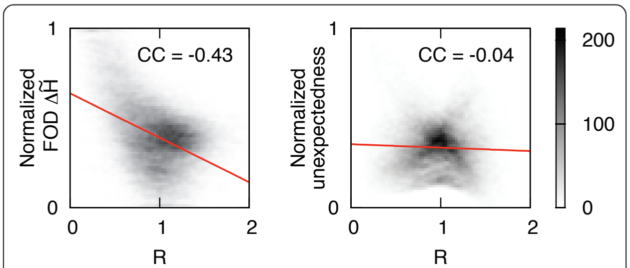


Figure 8 Dependence of $\Delta\tilde{H}$ and unexpectedness on reduced central distances. Two normalized global estimators of hydrophobic excess or deficiency, $\Delta\tilde{H}$ (according to the fuzzy oil drop model) and the unexpectedness, vs. reduced distance from the geometric center of a protein, *R*. Linear fits and correlation coefficients (CC) for 2D histograms were calculated for all atoms of unbound proteins from the LB_{48} test set.

description of hydrophobic preferences between (coarse-grained) scales of hydrophobicity and (fine-grained) residue-residue contact matrices, where more specific local effects such as homophilic, counterion or phenyl rings interactions can be expressed explicitly [74]. It has been already demonstrated that reduced representations and global geometric potentials are capable of a quantitative description of protein-ligand binding sites [75,76].

The adopted view concentrates on the characterization of proteins not assuming any specific chemical properties of ligands. Although based on a statistical model parametrized assuming spherical shapes of proteins (resembling the assumption behind the generalized Born solvation model), the method works well for moderately aspherical macromolecules, allowing for not only descriptive but also predictive applications. We do not incorporate into the identification method any additional features, such as the solvent accessible area or evolutionary conservation; the direct distance to the centroid was used only for the ranking in order to enable fair comparison with the FOD method; our measure is assigned homogeneously and isotropically in the whole protein volume, thus allowing for the examination of the predictive potential of the sole unexpectedness.

Favorable outcomes of our approach, especially when applied to enzymatic active sites, can be explained by analyzing the consequences of the requirement of the precise and resolute positioning of a ligand (as the prerequisite for chemical specificity), which can be best fulfilled by the creation of a binding pocket [77]. The burial of (still accessible) charged amino acids or the exposure of (partially unburied) conjugated aromatic ones, which are essential from the point of view of the mechanisms of the catalytic reactions, are not commensurate with their general expected radial positions in the bulk protein body. Frequently, despite their indented locations, pocket residues cannot be predominantly apolar as well, because of the need for the presence of bound water molecules assisting the catalysis (involved in, e.g., nucleophilic attack).

The most unexpected atoms are usually found in the deep-set parts of the pockets. The atomic depth has been found to be correlated with residue conservation [78,79] (more conserved amino acids create more contacts), which provides the explanation for the overlap between the sets of unexpected and conserved residues. It has been found, based on electrostatics, that functional sites comprise the most destabilizing residues [18]. Similarly, the unexpected amino acids are those introducing a local hydrophobic mismatch, plausibly counterbalanced by the formation of salt bridges and hydrogen bonding. The relation of the unexpectedness to the electrostatics is not, however, as simple as in the case of the conservation: buried charged residues can be encountered occasionally. It has been also demonstrated that electrostatic and hydrophobic interactions may compete [80]. This interplay is important with respect to the desolvation energy. The ease of desolvation is strongly predictive of protein-binding interfaces [29] and influences intricately ligand binding affinities [81]. As the hydrophobic interactions are dominant at protein interfaces [82], indicated scattered residues at the surface likely coincide with the view of the small fraction of hot-spots, which account for the majority of the binding energy [83].

Our approach yielded sets of parameters for every atom in an amino acid of a given type that is similar to the construction of a hydrophobicity scale, because the amount of information needed to characterize a protein is linearly proportional to the length of its sequence. The introduction of information-theoretic interpretation of hydrophobicity distributions may lead to valuable insights [84]. One result of the meeting of hydrophobicity and information theory, especially noteworthy in this context, supports our approach by demonstrating improvements in contact potentials tailored to the compositional properties of the sequences of interest [85].

The "mixture model" used in Equation 3 may be tuned via the expectation-maximization procedure to better fit the idealized distribution of the mass in individual proteins. However, we observed no improvement in the performance of the predictions for tuned forms, probably due to the already balanced composition of hydrophobic and polar amino acids in proteins selected by nature [86]. In this view, it would be interesting to check whether sequences of disordered or unfoldable structures give "mixture models" that deviate significantly from compact atomic distributions. It seems to be possible to apply the method from the smoothed surface towards the protein interior to some depth, and in this way cover proteins of more irregular shapes, consequently surpassing the most severe limitation of the approach. The attempt would require, however, the inquiry into the structure of hydrophobic cores in elongated or bent proteins.

The method is expected to be applicable for the functional annotation of low resolution structures, e.g., those resulting from mature homology modeling pipelines. Crude estimates of unexpectedness may be advantageous over computational geometry-based methods requiring precise atomic coordinates of active sites, where residues or even whole loops undergo significant displacements, not obeying the classic lock-and-key model [87].

Conclusion

We present an approach that captures orientational propensities of amino acids in globular proteins and offers a balanced description of their hydrophobic preferences. The description is created at the granularity of individual (amino acid-dependent types of) atoms but does not enumerate explicitly all possible interactions between them.

The approach is useful for the construction of a generic method that quantifies the unexpectedness of occurrences of individual atoms in a given distance from the geometric center of a protein. It turns out that the characteristics can be applied to the recognition of binding sites of both small ligands (enzymatic active sites) and other proteins (protein-protein interfaces).

Additional material

Additional file 1: Protein chains in the learning set.

Additional file 2: Geometric characteristics of the learning set and their dependencies.

Additional file 3: The plot of the probability density function used in this work.

Additional file 4: Parameters of atomic distributions.

Additional file 5: Probability densities of C^{α} and distal side chain atoms of 20 amino acids.

Additional file 6: Probability densities of C^{α} and distal side chain atoms of Cys.

Additional file 7: Details on the efficiency of the SurpResi applied to the LB_{48} test set.

Acknowledgements
The author would like to thank prof. I. Roterman for reading a preliminary version of the manuscript and dr. K. Prymula for discussions. A computational grant from the Academic Computer Center (ACK) CYFRONET AGH (MNiSW/IBM_BC_HS21/UJ/049/2009) is acknowledged.

Author details
[1]Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, ul. Reymonta 4, 30-059 Krakow, Poland. [2]Institute of Fundamental Technological Research, Polish Academy of Sciences, ul. Pawińskiego 5B, 02-106 Warsaw, Poland.

Authors' contributions
MK conceived of the study, implemented the method, carried out computations, analyzed results and wrote the manuscript.

Received: 22 May 2011 Accepted: 18 September 2011
Published: 18 September 2011

References

1. Li YY, Hou TJ, Goddard WA: **Computational modeling of structure-function of G protein-coupled receptors with applications for drug design.** *Curr Med Chem* 2010, **17(12)**:1167-80.
2. Fiorucci S, Zacharias M: **Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT.** *Proteins* 2010, **78(15)**:3131-9.
3. Seffernick JL, de Souza ML, Sadowsky MJ, Wackett LP: **Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different.** *J Bacteriol* 2001, **183(8)**:2405-10.
4. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA: **PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins.** *Nucleic Acids Res* 2004, W549-54.

5. Jambon M, Imberty A, Deléage G, Geourjon C: **A new bioinformatic approach to detect common 3D sites in protein structures.** *Proteins* 2003, **52(2)**:137-45.
6. Doppelt-Azeroual O, Delfaud F, Moriaud F, de Brevern AG: **Fast and automated functional classification with MED-SuMo: an application on purinebinding proteins.** *Protein Sci* 2010, **19(4)**:847-67.
7. Brylinski M, Skolnick J: **A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.** *Proc Natl Acad Sci USA* 2008, **105**:129-34.
8. Thangudu RR, Tyagi M, Shoemaker BA, Bryant SH, Panchenko AR, Madej T: **Knowledge-based annotation of small molecule binding sites in proteins.** *BMC Bioinformatics* 2010, **11**:365.
9. Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.** *J Mol Graph* 1995, **13(5)**:323-30, 307-8.
10. Brady GP Jr, Stouten PF: **Fast prediction and visualization of protein binding pockets with PASS.** *J Comput Aided Mol Des* 2000, **14(4)**:383-401.
11. Levitt DG, Banaszak LJ: **POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids.** *J Mol Graph* 1992, **10(4)**:229-34.
12. Hendlich M, Rippmann F, Barnickel G: **LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.** *J Mol Graph Model* 1997, **15(6)**:359-63, 389.
13. Weisel M, Proschak E, Schneider G: **PocketPicker: analysis of ligand binding-sites with shape descriptors.** *Chem Cent J* 2007, **1**:7.
14. Liang J, Edelsbrunner H, Woodward C: **Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.** *Protein Sci* 1998, **7(9)**:1884-97.
15. Le Guilloux V, Schmidtke P, Tuffery P: **Fpocket: an open source platform for ligand pocket detection.** *BMC Bioinformatics* 2009, **10**:168.
16. Coleman RG, Sharp KA: **Protein pockets: inventory, shape, and comparison.** *J Chem Inf Model* 2010, **50(4)**:589-603.
17. Yuan Z, Zhao J, Wang ZX: **Flexibility analysis of enzyme active sites by crystallographic temperature factors.** *Protein Eng* 2003, **16(2)**:109-14.
18. Elcock AH: **Prediction of functionally important residues based solely on the computed energetics of protein structure.** *J Mol Biol* 2001, **312(4)**:885-96.
19. Bate P, Warwicker J: **Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods.** *J Mol Biol* 2004, **340(2)**:263-76.
20. Laurie ATR, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.** *Bioinformatics* 2005, **21(9)**:1908-16.
21. Brylinski M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, Konieczny L, Roterman I: **Prediction of functional sites based on the fuzzy oil drop model.** *PLoS Comput Biol* 2007, **3(5)**:e94.
22. Oda A, Yamaotsu N, Hirono S: **Evaluation of the searching abilities of HBOP and HBSITE for binding pocket detection.** *J Comput Chem* 2009, **30(16)**:2728-37.
23. Bagley SC, Altman RB: **Characterizing the microenvironment surrounding protein sites.** *Protein Sci* 1995, **4(4)**:622-35.
24. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis.** *J Mol Biol* 1997, **272**:133-43.
25. Ondrechen MJ, Clifton JG, Ringe D: **THEMATICS: a simple computational predictor of enzyme function from structure.** *Proc Natl Acad Sci USA* 2001, **98(22)**:12473-8.
26. Bordner AJ: **Predicting small ligand binding sites in proteins using backbone structure.** *Bioinformatics* 2008, **24(24)**:2865-71.
27. Cilia E, Passerini A: **Automatic prediction of catalytic residues by modeling residue structural neighborhood.** *BMC Bioinformatics* 2010, **11**:115.
28. Panjkovich A, Daura X: **Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery.** *BMC Struct Biol* 2010, **10**:9.
29. Burgoyne NJ, Jackson RM: **Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces.** *Bioinformatics* 2006, **22(11)**:1335-42.
30. Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ: **Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties.** *PLoS Comput Biol* 2009, **5**:e1000266.
31. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA: **Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.** *PLoS Comput Biol* 2009, **5(12)**:e1000585.
32. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6**:19.
33. Bray T, Chan P, Bougouffa S, Greaves R, Doig AJ, Warwicker J: **SitesIdentify: a protein functional site prediction tool.** *BMC Bioinformatics* 2009, **10**:379.
34. Laskowski RA, Watson JD, Thornton JM: **ProFunc: a server for predicting protein function from 3D structure.** *Nucleic Acids Res* 2005, W89-93.
35. Huang B: **MetaPocket: a meta approach to improve protein ligand binding site prediction.** *OMICS* 2009, **13(4)**:325-30.
36. Brylinski M, Kochańczyk M, Konieczny L, Roterman I: **Sequence-structure-function relation characterized in silico.** *In Silico Biol* 2006, **6(6)**:589-600.
37. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272**:121-32.
38. Konieczny L, Brylinski M, Roterman I: **Gauss-function-based model of hydrophobicity density in proteins.** *In Silico Biol* 2006, **6(1-2)**:15-22.
39. Gomes ALC, de Rezende JR, Pereira de Araújo AF, Shakhnovich EI: **Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions.** *Proteins* 2007, **66(2)**:304-20.
40. Kauzmann W: **Some factors in the interpretation of protein denaturation.** *Adv Protein Chem* 1959, **14**:1-63.
41. Richards FM, Lim WA: **An analysis of packing in the protein folding problem.** *Q Rev Biophys* 1993, **26(4)**:423-98.
42. Dill KA: **Dominant forces in protein folding.** *Biochemistry* 1990, **29(31)**:7133-55.
43. Rackovsky S, Scheraga HA: **Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins.** *Proc Natl Acad Sci USA* 1977, **74(12)**:5248-51.
44. Jha AN, Vishveshwara S, Banavar JR: **Amino acid interaction preferences in proteins.** *Protein Sci* 2010, **19(3)**:603-16.
45. Nishikawa K, Ooi T: **Correlation of the amino acid composition of a protein to its structural and biological characters.** *J Biochem* 1982, **91(5)**:1821-4.
46. Taguchi Yh, Gromiha MM: **Application of amino acid occurrence for discriminating different folding types of globular proteins.** *BMC Bioinformatics* 2007, **8**:404.
47. Ma BG, Chen LL, Zhang HY: **What determines protein folding type? An investigation of intrinsic structural properties and its implications for understanding folding mechanisms.** *J Mol Biol* 2007, **370(3)**:439-48.
48. Rackovsky S: **Global characteristics of protein sequences and their implications.** *Proc Natl Acad Sci USA* 2010, **107(19)**:8623-6.
49. Roy S, Martinez D, Platero H, Lane T, Werner-Washburne M: **Exploiting amino acid composition for predicting protein-protein interactions.** *PLoS One* 2009, **4(11)**:e7813.
50. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-42.
51. Baumgärtner A: **Shapes of flexible vesicles at constant volume.** *J Chem Phys* 1993, **98**:7496-7501.
52. Galzitskaya OV, Bogatyreva NS, Ivankov DN: **Compactness determines protein folding type.** *J Bioinform Comput Biol* 2008, **6(4)**:667-80.
53. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247(4)**:536-40.
54. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH - a hierarchic classification of protein domain structures.** *Structure* 1997, **5(8)**:1093-108.
55. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases.** *Bioinformatics* 2001, **17(3)**:282-3.
56. Brylinski M, Kochanczyk M, Broniatowska E, Roterman I: **Localization of ligand binding site in proteins identified in silico.** *J Mol Model* 2007, **13(6-7)**:665-75.
57. Arteca GA: **Scaling behavior of some molecular shape descriptors of polymer chains and protein backbones.** *Phys Rev E* 1994, **49(3)**:2417-2428.
58. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, D129-33.

59. Sanner MF, Olson AJ, Spehner JC: **Reduced surface: an efficient way to compute molecular surfaces.** *Biopolymers* 1996, **38(3)**:305-20.
60. Hwang H, Vreven T, Janin J, Weng Z: **Protein-protein docking benchmark version 4.0.** *Proteins* 2010, **78(15)**:3111-4.
61. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z: **Integrating statistical pair potentials into protein complex prediction.** *Proteins* 2007, **69(3)**:511-20.
62. Pierce B, Weng Z: **ZRANK: reranking protein docking predictions with an optimized energy function.** *Proteins* 2007, **67(4)**:1078-86.
63. Li L, Guo D, Huang Y, Liu S, Xiao Y: **ASPDock: protein-protein docking algorithm using atomic solvation parameters model.** *BMC Bioinformatics* 2011, **12**:36.
64. Gabb HA, Jackson RM, Sternberg MJ: **Modelling protein docking using shape complementarity, electrostatics and biochemical information.** *J Mol Biol* 1997, **272**:106-20.
65. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z: **Protein-protein docking benchmark version 3.0.** *Proteins* 2008, **73(3)**:705-9.
66. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N: **ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information.** *Bioinformatics* 2003, **19**::163-4.
67. Eisenberg D, Weiss RM, Terwilliger TC: **The hydrophobic moment detects periodicity in protein hydrophobicity.** *Proc Natl Acad Sci USA* 1984, **81**:140-4.
68. Wu S, Liu T, Altman RB: **Identification of recurring protein structure microenvironments and discovery of novel functional sites around CYS residues.** *BMC Struct Biol* 2010, **10**:4.
69. Marino SM, Gladyshev VN: **Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces.** *J Mol Biol* 2010, **404(5)**:902-16.
70. Klotz IM: **Comparison of molecular structures of proteins: helix content; distribution of apolar residues.** *Arch Biochem Biophys* 1970, **138(2)**:704-6.
71. Lins L, Thomas A, Brasseur R: **Analysis of accessible surface of residues in proteins.** *Protein Sci* 2003, **12(7)**:1406-17.
72. Meirovitch H, Rackovsky S, Scheraga HA: **Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids.** *Macromolecules* 1980, **13(6)**:1398-1405.
73. Ben-Shimon A, Eisenstein M: **Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces.** *J Mol Biol* 2005, **351(2)**:309-26.
74. Singer MS, Vriend G, Bywater RP: **Prediction of protein residue contacts with a PDB-derived likelihood matrix.** *Protein Eng* 2002, **15(9)**:721-5.
75. Xie L, Bourne PE: **A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites.** *BMC Bioinformatics* 2007, **8(Suppl 4)**:S9.
76. Feldman HJ, Labute P: **Pocket similarity: are alpha carbons enough?** *J Chem Inf Model* 2010, **50(8)**:1466-75.
77. Campbell SJ, Gold ND, Jackson RM, Westhead DR: **Ligand binding: functional site location, similarity and docking.** *Curr Opin Struct Biol* 2003, **13(3)**:389-95.
78. Godzik A, Sander C: **Conservation of residue interactions in a family of Ca-binding proteins.** *Protein Eng* 1989, **2(8)**:589-96.
79. Pintar A, Carugo O, Pongor S: **Atom depth in protein structure and function.** *Trends Biochem Sci* 2003, **28(11)**:593-7.
80. Wang L, Friesner RA, Berne BJ: **Competition of electrostatic and hydrophobic interactions between small hydrophobes and model enclosures.** *J Phys Chem B* 2010, **114(21)**:7294-301.
81. Wang L, Berne BJ, Friesner RA: **Ligand binding to protein-binding pockets with wet and dry regions.** *Proc Natl Acad Sci USA* 2011, **108(4)**:1326-30.
82. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Natl Acad Sci USA* 1996, **93**:13-20.
83. Tuncbag N, Gursoy A, Keskin O: **Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy.** *Bioinformatics* 2009, **25(12)**:1513-20.
84. Pereira de Araujo AF, Onuchic JN: **A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins.** *Proc Natl Acad Sci USA* 2009, **106(45)**:19001-4.
85. Solis AD, Rackovsky SR: **Information-theoretic analysis of the reference state in contact potentials used for protein structure prediction.** *Proteins* 2010, **78(6)**:1382-97.
86. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Principal eigenvector of contact matrices and hydrophobicity profiles in proteins.** *Proteins* 2005, **58**:22-30.
87. Schmidt A, Lamzin VS: **Internal motion in protein crystal structures.** *Protein Sci* 2010, **19(5)**:944-53.
88. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH: **Hydrophobicity of amino acid residues in globular proteins.** *Science* 1985, **229(4716)**:834-8.
89. Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA: **Prediction of protein surface accessibility with information theory.** *Proteins* 2001, **42(4)**:452-9.
90. Biou V, Gibrat JF, Levin JM, Robson B, Garnier J: **Secondary structure prediction: combination of three different methods.** *Protein Eng* 1988, **2(3)**:185-91.
91. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C: **Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins.** *J Mol Biol* 1987, **195(3)**:659-85.
92. Guy HR: **Amino acid side-chain partition energies and distribution of residues in soluble proteins.** *Biophys J* 1985, **47**:61-70.
93. Wilce MCJ, Aguilar MI, Hearn MTW: **Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides.** *Analytical Chemistry* 1995, **67(7)**:1210-1219.

doi:10.1186/1472-6807-11-34
Cite this article as: Kochańczyk: Prediction of functionally important residues in globular proteins from unusual central distances of amino acids. *BMC Structural Biology* 2011 **11**:34.

Supplementary materials

for

Prediction of functionally important residues in globular proteins from unusual central distances of amino acids

Marek Kochańczyk[1,2,*]

[1] *Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, ul. Reymonta 4, 30-059 Krakow, Poland*

[2] *Institute of Fundamental Technological Research, Polish Academy of Sciences, ul. Pawińskiego, 02-106 Warsaw, Poland*

[*] *E-mail:* `marek.kochanczyk@uj.edu.pl`

2011

Table S1. PDB ids of 775 globular chains in the non-redundant learning set derived in this work.

1c3d_A	1wka_A	1kcq_A	2p3k_A	1d7p_M	1gv8_A	2vw5_A	1czs_A	1b1c_A	1eur_A	1ea5_A
1knb_A	1t1i_A	1thg_A	1srv_A	1gz7_A	1h49_A	2ri9_A	1m21_B	1myr_A	1wdp_A	2f6d_A
1nxc_A	1s95_A	2j75_B	1wcg_B	1hxn_A	1cge_A	1kex_A	1h10_A	1uuq_A	1ug6_A	2v3r_A
1ksc_A	1h46_X	1jdw_A	6cp4_A	1bn8_A	1h4p_B	1xwt_A	1ojj_B	1eqc_A	2fpv_A	1xnc_A
1t64_B	1r3r_A	1enu_A	1uqy_A	1dim_A	1edg_A	1r87_A	2d5j_A	1c3p_A	2aba_A	2d8l_A
966c_A	1f2j_A	3eau_A	1qjw_B	1eyw_A	2hu6_A	1xx2_A	1ocj_A	1g01_A	1vfl_A	1cem_A
1jta_A	1rgy_A	1ke4_B	1idk_A	1w3h_B	1qcx_A	1h6l_A	1xkn_A	1qaz_A	1a4m_A	1pw8_A
1w0h_A	1chd_A	1ezw_A	1fhw_A	1xyz_B	1lqa_B	2o0m_A	1qnp_A	1y65_A	1onr_B	1r66_A
1xfk_A	1j8t_A	1fob_A	1gxm_B	1hjs_A	1qjf_A	1n82_B	1ry8_A	1rhc_A	1g7f_A	2d2j_A
1jk7_A	1o3y_A	1vbr_B	2ghs_A	1mrq_A	1ds0_A	2gvv_A	1zpg_C	1v71_A	1lzl_A	1q5m_B
1rqj_A	2pll_B	1ppo_A	4lip_D	2iki_A	1gyh_A	1zgk_A	1tca_A	1qwk_A	1gmy_A	1frb_A
1l9x_C	1fhd_A	1up0_A	2hxz_B	1wl7_A	3c9e_A	1mlw_A	1v0k_A	1ute_A	1pyf_A	2cyg_A
1gx4_A	2had_A	2ixt_A	1rkd_A	1hdu_A	5a3h_A	1lyv_A	2p4o_A	1h1n_A	2p41_A	1cqw_A
1gok_A	1dqy_A	1g4h_A	1bqc_A	1nq6_A	1lok_A	1cnv_A	1q0z_A	1wb4_B	1ltu_A	1jln_A
1hq0_A	1ls6_A	2gmn_A	1nnh_A	1mtz_A	1uv4_A	1u5h_A	1bf6_A	1n57_A	1eok_A	1nar_A
1b8o_A	2i47_A	1o4y_A	1pzt_A	1tml_A	1fqg_A	1q7f_A	1f0n_A	1qtw_A	1tkj_A	1ukb_A
1y21_A	1a2q_A	1r88_B	2pkc_A	1va4_A	1hl7_B	1thm_A	1umz_B	1i6n_A	2gu5_B	1qqf_A
1brt_A	1deu_A	1wma_A	1om0_A	1ja9_A	1a8q_A	1llo_A	1a8s_A	1dyp_A	1ak0_A	1jov_A
3b68_A	2nv6_A	1sml_A	1st3_A	1tib_A	3tgl_A	1jjf_A	1ako_A	1pa7_A	1j6o_A	1j02_A
1fsf_A	2gnp_A	1tys_A	2a0n_A	1lug_A	1qhv_A	2cdd_A	1a28_A	1thf_D	1ny1_B	2pbl_B
1jfr_B	1vic_A	1b0u_A	1o0x_A	2f1n_A	1sep_A	1uza_A	1xfj_A	1v9e_A	1p1x_A	2nmx_B
1k77_A	1txo_A	1rv9_A	1zzm_A	1oha_A	1m33_A	3c70_A	2bwa_A	1xa1_B	1ax2_A	1h70_A
1caq_A	1jyk_A	1b2l_A	1vc4_B	2q46_A	1x06_A	1gqn_A	1km3_A	2gpu_A	1zk0_A	1n55_A
1g24_A	1qwg_A	2o34_A	1e58_A	1lyx_A	1qrl_A	1jzt_A	1pbt_A	1udh_A	1rw7_A	1mve_A
1i9t_A	1qo2_A	1vyb_B	2qfo_B	1o1y_A	1ini_A	1nxd_3	1din_A	1mvq_A	1fx2_A	1jg4_A
2nlr_A	1xjz_A	1k4l_A	1dex_A	1wab_A	1fj2_B	1fx4_A	1eug_A	1jjt_A	1dxk_A	1g61_B
1g6c_A	1nfp_A	1kdt_A	2abw_B	2pof_A	1gxy_A	1i1n_A	1upi_A	1dak_A	1uu6_A	1uai_A
1u9c_A	2cl5_B	2hxm_A	3b5e_B	1njs_B	1oq1_C	1okb_A	1oa4_A	1k7j_A	1aec_A	1q7r_A
1uol_B	1q0u_A	1oa2_F	2o2x_A	1fva_A	1l8b_B	1nn1_A	3c7i_A	1jfx_A	1txl_A	1wnx_B
1g3u_A	2ayh_A	1agy_A	1r55_A	1l8f_A	1gbg_A	1ijb_A	1cpn_A	1pt6_B	4tmk_A	4eng_A
3gar_A	1v77_A	1ppn_A	1hbp_A	1ff3_C	1yzq_A	1hjz_B	1h4h_D	1p3u_A	1aun_A	1dix_A
1o0e_A	1pzs_A	1d4o_A	1h2e_A	1vg8_A	2blu_A	1nf8_A	1cjw_A	1qoz_B	1bs9_A	2cd2_A
1he4_A	1u8y_B	1ui0_A	1lhu_A	1jm1_A	1vk2_A	1kuf_A	1ukz_A	1nwa_A	1nd1_A	1h4e_A
1vp8_A	2cdn_A	1iqq_A	1wc9_A	1f5j_A	2gf0_A	2i6g_B	1e87_A	1tf1_B	1z06_A	1p5f_A
1bsw_A	1m55_B	1j1f_A	1ido_A	1xnk_B	1rie_A	1atz_B	1kmq_A	2atv_A	1ioo_A	1g5t_A
1el4_A	1x3s_A	2nr7_A	1jf0_A	1qf9_A	1tc5_D	1lm4_A	1yna_A	1dus_A	1j54_A	1pvx_A
2pth_A	1ia1_B	1mh1_A	1oqv_A	1uxo_A	2isb_A	1uhh_B	2o7n_A	1sl8_A	1oix_A	1pl3_B
2dfb_A	1hzt_A	8dfr_A	1ihc_A	1qra_A	1i8a_A	1eq6_A	1vkf_C	1bsz_B	1koe_A	1mvt_A
1eyl_A	1beh_A	1jfu_A	1l1q_A	2if6_A	1gbs_A	1pmh_X	1u17_B	1epz_A	1r8n_A	1ywd_A
2dfn_A	2nn5_A	1lqy_A	1nxj_B	1x1r_A	1kn3_A	1ky2_A	1h0p_A	2i6c_A	2fn4_A	2qxu_H
1mr3_F	1fzq_A	1n5n_B	1tiq_A	2eu7_X	1j83_B	1vjf_A	1rxd_B	1eiz_A	1im5_A	1i06_A
2ot9_A	1jwq_A	1oh4_A	1yzl_A	1iko_P	1znk_A	1a58_A	2hia_A	1ghe_B	2a2n_C	1gwy_B
1i6t_A	1qfv_B	1vhs_A	1wba_A	1vhh_A	1sl5_A	1vi4_A	1cv8_A	1euj_A	1qmy_A	2cyh_A
2fko_A	1dyw_A	1h4o_C	1zmf_A	2fcr_A	1mmq_A	2ery_B	2c8s_A	2ow9_A	1jhj_A	1n6n_A
2bem_C	1ek0_A	1vi3_A	1yvd_A	1ofv_A	1obo_B	1rm8_A	1mug_A	2cua_A	1f3z_A	1z4r_A
1z2a_A	1od3_A	1ddw_A	1uuy_A	1kao_A	2nvh_A	2gkp_A	1xo7_A	1qst_A	1vai_A	1nyk_A
2bit_X	1zp5_A	1g81_A	1nrz_D	1dly_A	1sen_A	1mz4_A	1n08_B	1uz2_X	1ist_B	1gpr_A
3dfr_A	1tp9_C	1htw_A	1d2a_A	2ijq_A	2icg_A	1mxi_A	1fm4_A	1mfm_A	1ra8_A	1dg7_A
1q0n_A	1xdf_B	1edu_A	2hbo_A	1dg9_A	1jyh_A	1e00_A	2fqt_A	1elk_A	1kva_A	1kng_A
1bj7_A	1gy1_A	1npk_A	1bfg_A	1ab0_A	2oeb_A	1o1x_A	1icx_A	1m16_B	1gui_A	2spo_A
1oj6_D	1yaz_A	1o7u_A	1md6_A	1emy_A	1mno_B	2nsr_A	1gdj_A	1gwm_A	2ob5_A	1l1d_B
1lic_A	1q0e_A	1fg4_A	1id0_A	1oal_A	1w1g_A	2i8g_A	1e5p_B	1st9_A	1stn_A	1oz9_A
1zzo_A	1nb9_A	1akt_A	1t2w_C	1mba_A	1h97_B	1jf4_A	1o4w_A	1kjl_A	1it2_A	2oyn_A
2hd9_A	1uy3_A	1p90_A	1at0_A	2d59_A	1lit_A	1q1u_A	1j7g_A	1b20_A	1tzx_B	1w9t_A
1hdk_A	1gz2_A	1iuk_A	1ov8_B	1rfs_A	1fvx_A	1ktg_A	1lhi_A	1jer_A	2fs6_B	3bzp_A
1c1f_A	1xs0_A	1eca_A	1o13_A	1mvo_A	1moy_A	1pdo_A	1lu4_A	2aif_A	1is6_A	3gal_A
1vyf_A	1p0z_A	1dqg_A	1e29_A	1r9h_A	1fsj_B	1opb_C	1uc7_B	1tu9_A	2fuf_A	1o8v_A
2ia7_A	1kqw_A	1zwz_A	1bea_A	1c52_A	1uxx_X	2ohw_B	1c7k_A	1srr_A	1icm_A	1hmt_A
1mdc_A	1lju_A	3nul_A	1jb3_A	1mai_A	1wna_A	1cc3_A	1mc9_A	1r29_A	2czw_A	1chn_A
1cuo_A	2ccw_A	1ow4_B	1u79_A	1cot_A	1a4a_B	1ou8_B	1i3u_A	1u29_A	2bt6_A	1ijt_A
1tp6_A	2fi9_A	1doi_A	1ijx_A	1zes_A	1rzy_A	1dbw_B	2gte_B	1fao_A	1t1j_A	1ugu_A
1jug_A	1r26_A	1eaz_A	8paz_A	1cxc_A	2cw4_A	1oae_A	1zia_A	1m5t_A	1rtx_A	1f9m_A
2fc3_A	1v30_A	1c44_A	2trx_A	1hq8_A	1hxr_B	1upq_A	1pmy_A	1wou_A	1ufy_A	1f7l_A
1whi_A	1qto_A	1m9z_A	3b7c_A	2pl1_A	1buo_A	1ifr_A	2a9o_A	1tmy_A	1ikt_A	1ra4_A
1tq3_A	1o7i_B	1opc_A	2cyj_A	1h4y_B	1h8u_A	2fne_B	2byg_A	2od5_A	4fiv_A	1gou_B
1pz4_A	1dlw_A	1mg4_A	6fiv_A	1td0_D	1thx_A	2pyq_A	1rtu_A	1svy_A	2iay_A	1o4i_A
1n8v_B	1dw0_B	1ytc_A	3c2c_A	2q3w_A	1ccr_A	2o3f_C	1i7h_C	1qwx_B	1pva_A	1kr7_A
1rwy_A	1n9l_A	5pal_A	1a75_B	2dg3_A	1tuw_A	1b8r_A	1bkr_A	1rwy_B	1bu3_A	1irv_A
1kaf_B	1omd_A	1gn0_A	1i1j_B	1d3w_A	1co6_A	2q5b_B	1iib_B	1oqq_B	2r48_A	1erw_A
1t5k_B	1rms_A	1hrc_A	2fmb_A	1i0x_D	1ln4_A	5cyt_R	2h3l_B	1xmt_A	1h7m_A	1tsf_A
4vub_A	2bo1_A	1l8r_A	1o5u_A	1n3y_A						

Table S2. Parameters of the probability density function: $p(R;\tau)=\frac{A_\tau R^{\gamma_\tau}}{1+\exp(\beta_\tau(R-\mu_\tau))}$ derived from histograms of reduced central distances, R, collected for all types of heavy atoms of all types of amino acids as ocurring in globular proteins from the learning set (listed in the supplementary Table S1). Heavy atoms in side chains that are characterized by central distances distributed significantly different from central distances of C^{α} atoms (Kolmogorov-Smirnov tests with p-value < 0.000001) were marked with stars.

τ=(Aa,	Atom)	A_τ	μ_τ	β_τ	γ_τ
A	N	1.85954	1.14344	9.62103	2.01663
A	O	1.88563	1.14069	9.13446	2.1443
A	C	1.919	1.13577	9.5249	2.12155
A	C^{α}	1.86656	1.13408	8.70091	1.99614
· A	C^{β}	1.90715	1.10367	7.09949	1.9342
C_{SS}	N	2.42082	1.07441	8.54337	3.0098
C_{SS}	O	2.16193	1.11001	8.78373	2.88278
C_{SS}	C	2.22361	1.10463	9.01742	2.9227
C_{SS}	C^{α}	2.4535	1.06968	8.54859	2.96551
· C_{SS}	C^{β}	3.61427	0.969087	8.10373	3.65727
· C_{SS}	S^{γ}	4.87096	0.895388	7.75633	3.96847
C_{SH}	N	2.24242	1.01742	9.4792	1.49511
C_{SH}	O	2.03475	1.0501	9.11637	1.43533
C_{SH}	C	2.06666	1.04238	9.39009	1.41924
C_{SH}	C^{α}	2.51755	0.980198	8.93293	1.60296
· C_{SH}	C^{β}	2.69044	0.953012	8.91117	1.59813
· C_{SH}	S^{γ}	2.98544	0.92291	8.7473	1.68042
C	N	2.20509	1.0562	8.49479	1.85681
C	O	2.0752	1.09025	8.52757	1.80324
C	C	2.03371	1.08615	8.90352	1.78213
C	C^{α}	2.42883	1.02375	8.19802	1.95379
· C	C^{β}	2.77991	0.979176	7.9176	2.05535
· C	S^{γ}	3.06552	0.947826	7.78738	2.11068
D	N	1.85486	1.18094	11.289	3.27425
D	O	1.89458	1.1661	10.1346	3.40543
D	C	1.89924	1.17198	10.967	3.40343
D	C^{α}	1.74977	1.19626	11.1195	3.43601
* D	C^{β}	1.46201	1.25025	10.8146	3.22751
* D	C^{γ}	1.3105	1.285	11.0274	3.27626
* D	$O^{\delta 1,\delta 2}$	1.23955	1.30143	10.6261	3.23178
E	N	1.98557	1.15871	10.8956	3.64001
E	O	1.85071	1.17845	10.8713	3.46577
E	C	1.9308	1.16974	11.357	3.57748
E	C^{α}	1.77466	1.19266	11.4089	3.64965
* E	C^{β}	1.48677	1.24168	11.0907	3.54274
* E	C^{γ}	1.31626	1.27515	10.8993	3.57997
* E	C^{δ}	1.7199	1.2982	10.7479	3.88045
* E	$O^{\epsilon 1,\epsilon 2}$	1.098	1.31389	10.4551	3.89444
F	N	2.26286	1.07153	10.3102	1.80383
F	O	2.23484	1.06709	9.84234	1.93161
F	C	2.20673	1.07025	10.435	1.87167
F	C^{α}	2.33045	1.04534	10.1424	1.84731
· F	C^{β}	2.70951	0.994458	9.58106	1.94113
* F	C^{γ}	3.21173	0.938349	8.82322	2.0412
* F	$C^{\delta 1,\delta 2}$	3.43458	0.912336	8.25185	2.07026
* F	$C^{\epsilon 1,\epsilon 2}$	3.97823	0.861718	7.50868	2.15559
* F	C^{ζ}	4.07383	0.850231	7.34663	2.14493
G	N	1.41659	1.25703	10.7643	1.83934
G	O	1.52184	1.21896	8.89533	1.944456
G	C	1.51989	1.22519	9.73177	1.95465
G	C^{α}	1.36597	1.26867	9.86165	1.8015
H	N	1.79522	1.14741	10.5185	1.77569
H	O	1.64444	1.17279	10.3303	1.60987
H	C	1.71863	1.1582	10.4955	1.67697
H	C^{α}	1.74533	1.15543	10.1783	1.73945
· H	C^{β}	1.79077	1.14146	8.99532	1.79405
· H	C^{γ}	1.66554	1.16869	8.73623	1.75192
* H	$N^{\delta 1}$	1.55861	1.19585	8.51195	1.72934
* H	$C^{\delta 2}$	1.56118	1.1889	8.20466	1.66551
* H	$C^{\epsilon 1}$	1.38993	1.24323	8.33837	1.613
* H	$N^{\epsilon 2}$	1.3804	1.24305	8.17855	1.56715
I	N	2.28888	1.0443	10.0481	1.76938
I	O	1.97195	1.09031	10.7303	1.58805
I	C	2.0243	1.07838	10.7318	1.58331
I	C^{α}	2.45109	1.01476	9.4227	1.7797
· I	C^{β}	3.0675	0.940875	8.31378	1.96042
* I	$C^{\gamma 1}$	3.95679	0.875253	7.82796	2.22682
* I	$C^{\gamma 2}$	3.27028	0.908294	7.47578	1.97036
* I	$C^{\delta 1}$	4.95071	0.818597	7.43178	2.43311
K	N	2.73216	1.07917	10.8264	4.74762
K	O	2.10346	1.13886	10.4785	3.95028
K	C	2.33073	1.11818	11.0779	4.28649
K	C^{α}	2.39826	1.10889	11.0951	4.81372
* K	C^{β}	2.08623	1.13076	10.6215	5.00503
* K	C^{γ}	1.72442	1.17117	10.461	4.91766
* K	C^{δ}	1.41126	1.20814	10.3063	5.11234
* K	C^{ϵ}	1.15114	1.25083	10.2292	5.15416
* K	N^{ζ}	0.95605	1.29372	10.1785	5.12301
L	N	2.08193	1.09252	10.6923	1.84865
L	O	2.00599	1.10513	10.0332	1.85942
L	C	2.04411	1.09916	10.5414	1.84819
L	C^{α}	2.21541	1.06739	10.1249	1.87492
* L	C^{β}	2.56815	1.01039	8.90099	1.96028
* L	C^{γ}	3.08358	0.94331	7.82283	2.07451
* L	$C^{\delta 1,\delta 2}$	3.7841	0.877396	7.04436	2.26302
M	N	2.16412	1.07156	9.2319	1.88134
M	O	2.2221	1.06911	9.223	1.98419
M	C	2.19126	1.07248	9.55179	1.92684
M	C^{α}	2.3475	1.04142	8.86534	1.93347
· M	C^{β}	2.57036	0.998159	7.92453	1.93913
· M	C^{γ}	3.31598	0.903553	6.74281	2.13118
* M	S^{δ}	4.4474	0.80527	6.1702	2.29823
* M	C^{ϵ}	4.75168	0.773953	5.79667	2.34705
N	N	2.02814	1.14674	10.2247	3.1657
N	O	1.97925	1.14291	9.10539	3.08941
N	C	2.0731	1.13614	9.72506	3.21494
N	C^{α}	1.91178	1.16099	9.83299	3.18936
* N	C^{β}	1.60765	1.2129	9.53132	2.97622
* N	C^{γ}	1.39266	1.26559	9.86253	2.8308
* N	$O^{\delta 1}$	1.3314	1.2813	9.66778	2.75331
* N	$N^{\delta 2}$	1.28146	1.29503	9.68075	2.76466

Continued on the next page.

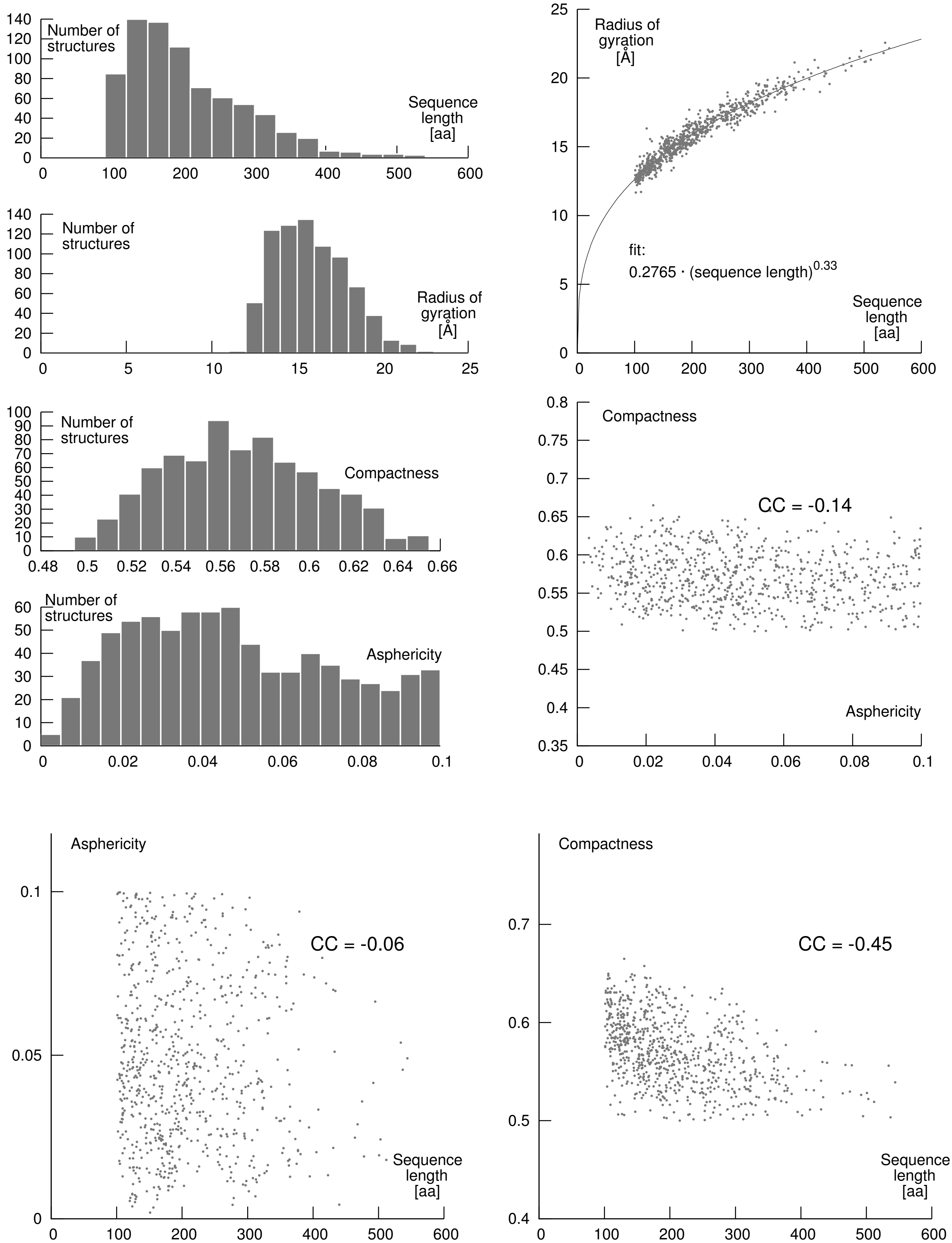
Table S2. – *Continued.*

τ=(Aa,	Atom)	A_τ	μ_τ	β_τ	γ_τ
P	N	1.74989	1.19768	11.1837	3.07058
P	O	1.73507	1.19726	10.6971	3.09794
P	C	1.72868	1.20225	11.3192	3.077
P	C^α	1.65503	1.21496	11.1319	2.9767
* P	C^β	1.39257	1.26921	10.0829	2.65189
* P	C^δ	1.52256	1.23685	10.1116	2.77894
* P	C^γ	1.34911	1.27895	9.66413	2.56784
Q	N	2.08951	1.14403	10.8973	3.28799
Q	O	1.91142	1.16703	10.6304	3.07799
Q	C	1.96307	1.16312	11.2172	3.12275
Q	C^α	1.90483	1.17245	11.1804	3.2515
* Q	C^β	1.66129	1.21203	10.922	3.14764
* Q	C^δ	1.34825	1.27416	10.5634	3.22498
* Q	C^γ	1.51229	1.23811	10.5453	3.20969
* Q	$O^{\epsilon 1}$	1.30044	1.28201	10.0559	3.22444
* Q	$N^{\epsilon 2}$	1.23968	1.30301	10.4475	3.13179
R	N	2.61297	1.08558	10.7821	3.58399
R	O	2.2283	1.1188	10.1269	3.20002
R	C	2.37103	1.10773	10.6765	3.35447
R	C^α	2.40342	1.10845	10.9546	3.5691
* R	C^β	2.14991	1.13602	10.7208	3.54819
* R	C^γ	2.03587	1.14804	10.4279	3.67619
* R	C^δ	1.75977	1.18286	9.88739	3.62322
* R	N^ϵ	1.72646	1.18439	9.83642	3.84888
* R	C^ζ	1.63791	1.19295	9.59623	3.99212
* R	$N^{\eta 1,\eta 2}$	1.54164	1.20664	9.31346	3.94206
S	N	1.65424	1.19857	9.83601	2.19989
S	O	1.64665	1.20153	9.61894	2.30612
S	C	1.66862	1.19903	9.96855	2.31836
S	C^α	1.53058	1.22762	9.73145	2.14097
* S	C^β	1.3328	1.28044	9.26341	1.94463
* S	O^γ	1.33149	1.28084	9.09863	2.01394
T	N	1.91911	1.1521	10.5342	2.36853
T	O	1.87189	1.15123	9.33962	2.308
T	C	1.90488	1.15021	9.89948	2.34845
T	C^α	1.78745	1.17279	10.157	2.26954
* T	C^β	1.56428	1.21837	9.83832	2.10378
* T	$O^{\gamma 1}$	1.53926	1.22632	9.78322	2.1719
* T	$C^{\gamma 2}$	1.40288	1.25504	9.2939	1.85524
V	N	2.14248	1.06128	9.61542	1.69647
V	O	1.95298	1.09021	9.79511	1.58572
V	C	2.01885	1.07492	9.627664	1.58672
V	C^α	2.31811	1.02622	8.74914	1.72555
· V	C^β	2.82571	0.959132	7.78537	1.91297
* V	$C^{\gamma 1,\gamma 2}$	3.4067	0.898944	7.10407	2.09198
W	N	2.16666	1.06852	9.68157	1.8162
W	O	2.52022	1.02418	8.45761	2.10325
W	C	2.60672	1.01789	8.88978	2.12404
W	C^α	2.48678	1.02519	9.18021	1.96748
· W	C^β	2.74637	0.993851	9.08366	2.03041
· W	C^γ	2.87076	0.992432	9.78947	2.12948
· W	$C^{\delta 1}$	2.39333	1.04045	9.45923	1.95456
* W	$C^{\delta 2}$	3.2157	0.97072	10.5317	2.25817
· W	$N^{\epsilon 1}$	2.47356	1.03365	9.33899	2.03229
· W	$C^{\epsilon 2}$	2.9114	0.991737	9.76106	2.18207
* W	$C^{\epsilon 3}$	3.66434	0.93326	10.0902	2.33315
· W	$C^{\zeta 2}$	2.87099	0.988483	9.15555	2.14396
* W	$C^{\zeta 3}$	3.60627	0.928426	9.44348	2.28163
· W	$C^{\eta 2}$	3.14229	0.959377	9.06995	2.16779
Y	N	2.45871	1.06558	10.5233	2.36149
Y	O	2.55032	1.0485	9.64581	2.38468
Y	C	2.57736	1.05001	10.2519	2.39466
Y	C^α	2.68635	1.0426	10.4916	2.47518
· Y	C^β	3.1316	1.00069	9.91091	2.65577
· Y	C^γ	2.96367	1.01834	10.1356	2.64251
· Y	$C^{\delta 1,\delta 2}$	2.72005	1.0364	9.81383	2.52674
* Y	$C^{\epsilon 1,\epsilon 2}$	2.35074	1.07434	9.54375	2.36893
* Y	C^ζ	2.20694	1.09349	9.5077	2.31764
* Y	O^η	1.8242	1.1511	9.06886	2.09818

Table S3. Geometrical and functional characteristics of structures from the non-redundant apo-holo set created by Hwang and Schroeder (*BMC Struct Biol* 6:19 (2006)) and performance of two binding site recognition methods for two cutoff distances. Geometrical descriptors, aspericity (Asph.) and compactness (Comp.), are reported for apoproteins. The list is ordered according to the increasing asphericity. When a method was unable to find a site, the rank is ∞. Enzyme class assignments according to the Catalytic Site Atlas (*Nucleic Acids Res* 32:D129-33 (2004)) version 2.2.12 (January 2010).

PDB and chain ID			Enzyme class or the non-substrate ligand	Geometry		FOD		SurpResi	
Apo	Holo —	Ligands		Asph.	Comp.	4 Å	6 Å	4 Å	6 Å
1nna A	1ivd A	FUL, ST1, NAG, MAN	3.2.1.18	0.0144	0.5617	1	1	1	1
2sil A	2sim A	DAN	3.2.1.18	0.0198	0.555	1	2	1	1
2ctb A	2ctc A	HFA	3.4.17.1	0.0236	0.5836	∞	2	1	1
5cpa A	7cpa A	FVF	3.4.17.1	0.0237	0.5744	∞	1	∞	1
1brq A	1rbp A	RTL	retinol	0.0237	0.5431	1	1	∞	2
1hxf H	1dwd H	MID	3.4.21.5	0.0241	0.536	∞	1	2	2
2cba A	2h4n A	AZM	4.2.1.1	0.025	0.5497	1	1	1	1
4ca2 A	1okm A	SAB	4.2.1.1	0.0251	0.5592	1	1	∞	1
1cge A	1hfc A	PLH	3.4.24.7	0.037	0.5682	1	1	1	1
1esa A	1inc A	ICL	3.4.21.36	0.0375	0.5595	∞	1	∞	1
1stn A	1snc A	THP	3.1.31.1	0.0378	0.56	1	1	1	1
1chg A	3gch C	OAC	3.4.21.1	0.0397	0.5645	∞	∞	∞	2
1ypi A	2ypi A	PGA	5.3.1.1	0.041	0.5401	1	1	1	1
3tms A	1bid A	FMT,UMP	2.1.1.45	0.0428	0.52	1	1	1	1
1ifb A	2ifb A	PLM	fatty acid	0.0446	0.6057	1	1	1	1
1qif A	1acj A	THA	3.1.1.7	0.0451	0.4984	∞	2	1	1
1ula A	1ulb A	GUN	2.4.2.1	0.0562	0.5017	∞	∞	∞	1
1ime A	1imb A	LIP	3.1.3.25	0.0588	0.5353	1	1	1	1
3ptn A	3ptb A	BEN	3.4.21.4	0.0602	0.598	1	1	1	1
2tga A	1mtw A	DX9	3.4.21.4	0.0617	0.6043	∞	∞	∞	2
1bya A	1byb A	GLC	3.2.1.2	0.0646	0.51	1	1	1	1
1krn A	2pk4 A	ACA	3.4.21.7	0.0764	0.6402	∞	∞	1	1
1phc A	1phd A	HEM,PIM	1.14.15.1	0.0795	0.508	1	1	4	4
5dfr A	4dfr A	MTX	1.5.1.3	0.0806	0.5479	1	1	1	1
1djb A	1blh A	FOS	3.5.2.6	0.0833	0.5586	1	1	2	1
2fbp B	1fbp B	AMP,F6P	3.1.3.11	0.0864	0.4873	1	2	1	1
1ahc A	1mrg A	ADN	3.2.2.22	0.0937	0.5505	∞	1	3	1
1pdy A	1pdz A	ACE,PGA	4.2.1.11	0.096	0.5315	2	1	2	1
2ctv A	5cna A	MMA	saccharide	0.0979	0.5612	∞	∞	∞	∞
1hsi A	1ida A	QND, PR0, PY2, PPL	3.4.23.47, 2.7.7.7, 3.1.26.13, 2.7.7.49	0.1121	0.5362	1	1	2	1
1bbs A	1rne A	NGA,C60	3.4.23.15	0.1159	0.5135	1	1	1	1
1hel A	1hew A	NAG	3.2.1.17	0.1204	0.6043	∞	1	5	5
3phv A	4phv A	VAC	2.7.7.49, 3.4.23.16, 2.7.7.7, 3.1.26.13	0.1219	0.5386	∞	∞	∞	1
6ins E	3mth A	MPB	benzoic acid ester	0.1259	0.5948	∞	∞	∞	∞
3app A	1apu E	IVA,STA,EHN	3.4.23.20	0.1274	0.5507	1	1	1	1
3lck A	1qpe A	PP2,PTR	2.7.10.2	0.1356	0.4987	∞	1	∞	2
1psn A	1pso E	IVA,STA	3.4.23.1	0.1417	0.5352	2	2	2	1
8rat A	1rob A	C2P	3.1.27.5	0.1427	0.5795	1	1	2	1
1a6u H	1a6w H	NIP	iodonitrophenylacetyl-aminocaproic acid	0.1446	0.6371	∞	∞	∞	2
1swb A	1stp A	BTN	biotin	0.145	0.542	1	1	∞	∞
1pts A	1srf A	MTB	azobenzoic acid	0.1553	0.5564	1	1	1	4
3p2p A	5p2p A	DHG	3.1.1.4	0.1667	0.5674	1	1	1	1
2rta A	1stp A	BTN	biotin	0.1703	0.5332	∞	1	∞	∞
8adh A	1cdo A	NAD	1.1.1.1	0.1795	0.5026	1	1	1	1
1l3f E	2tmn E	0FA	3.4.24.27	0.1966	0.573	∞	1	1	1
1npc A	1hyt A	DMS,BZS	3.4.24.28	0.2032	0.557	1	1	1	1
1gcg A	1gca A	GAL	aldohexose	0.2881	0.5325	1	1	4	2
1a4j B	1igj D	DGX	digoxin	0.4645	0.4737	∞	∞	∞	∞

Figure S1. Histograms and dependencies of several characteristics of the learning protein set (CC – correlation coefficient).



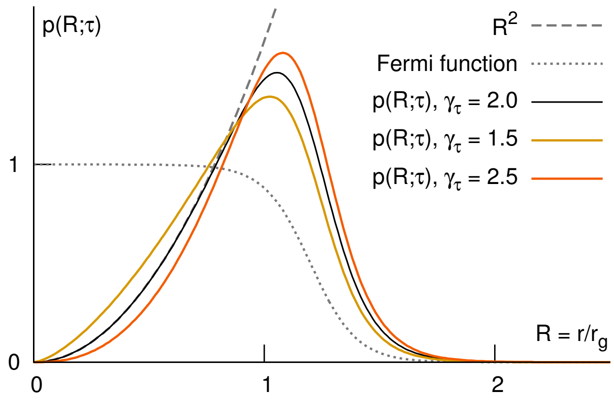


Figure S2. Probability density function used in this work:

$$p(R;\tau) = \frac{A_\tau\, R^{\gamma_\tau}}{1 + \exp(\beta_\tau(R - \mu_\tau))}$$

for example values of γ_τ. Parameters: $\mu_\tau = 1.2$, $\beta_\tau = 10$; values of A_τ are chosen accordingly to normalize distributions. In general, when $\gamma_\tau < 2$, the function better fits histograms of atomic central distances for hydrophobic amino acids; when $\gamma_\tau > 2$, it better fits histograms of atomic distances for hydrophilic residues. For $\gamma_\tau = 2$ the function adopts the simplest form of a special case ($\alpha_\tau = 1$) demonstrated by Gomes et al. (*Proteins* 66(2):304-20 (2007)). The location of the maximum is $\beta^{-1}\,(\gamma + W(\gamma \exp(\beta\mu - \gamma)))$, where W is the (Lambert's) omega function, and the mean value can be estimated by $-A\beta^{-(\gamma+2)}\,\Gamma(\gamma+2)\,\mathrm{Li}_{\gamma+2}\,(-\exp(\beta\mu))$, where Γ and Li are the (Euler's) gamma and (Jonquière's) polylogarithm functions.

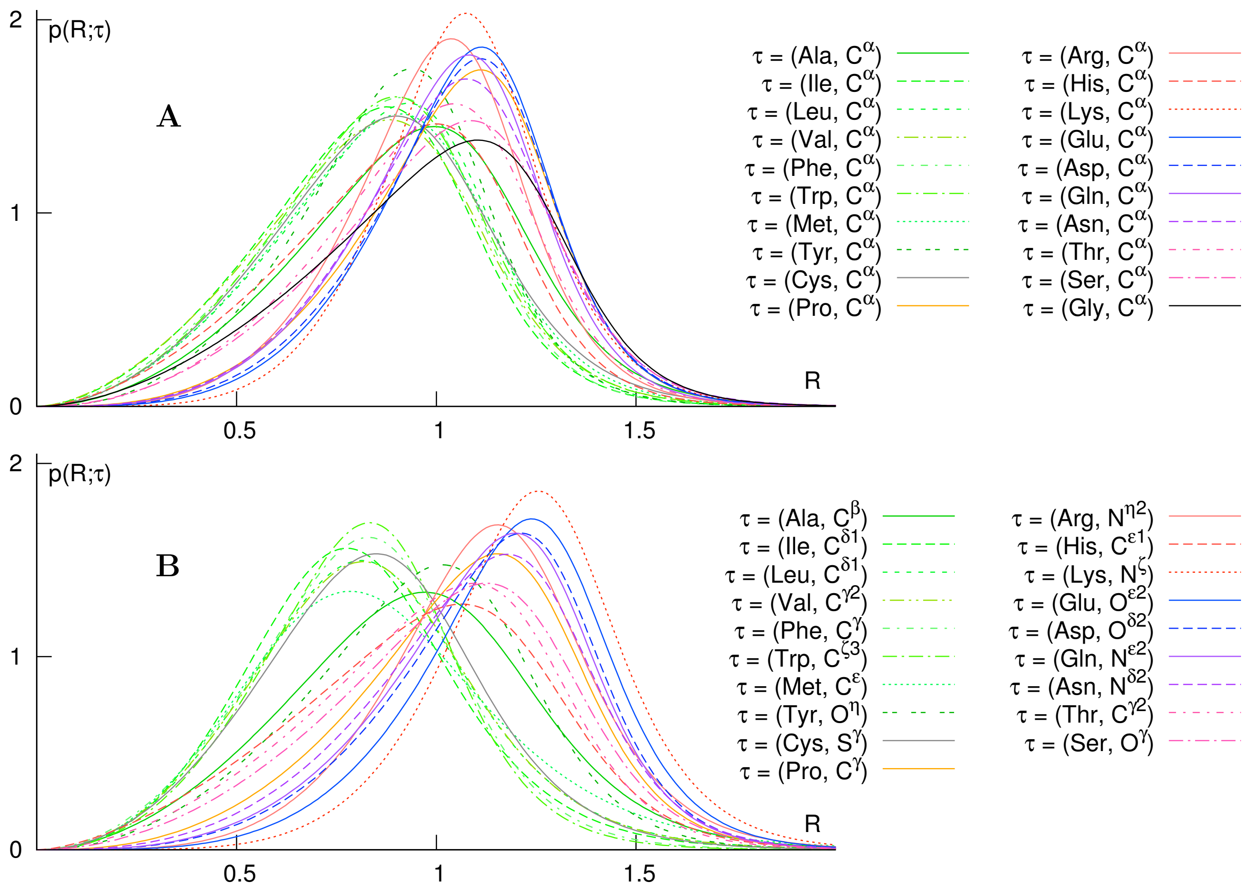


Figure S3. Distributions of central distances of Cα (**A**) and distal side chain atoms (**B**) of all amino acids. Curves for amino acids with hydrophobic side chains are green, polar charged – red and blue, polar uncharged – pink and violet.

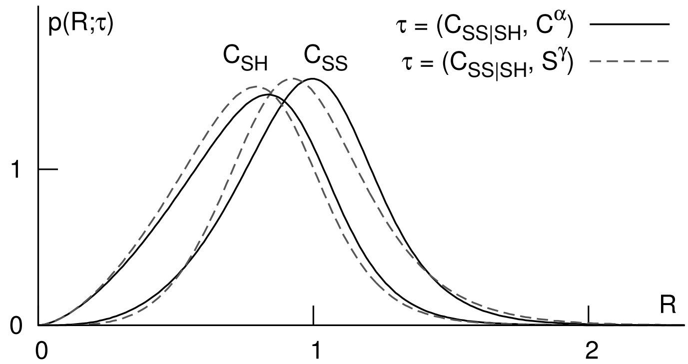


Figure S4. Probability densities of C^α and distal side chain atoms of Cys. Two cases are shown separately: Cys bridged (C_{SS}) and not bridged (C_{SH}) by disulfide bonds.

Supplementary references

[Gomes et al., 2007] Gomes, A. L. C., de Rezende, J. R., Pereira de Araújo, A. F., and Shakhnovich, E. I. (2007). Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. *Proteins*, 66(2):304–20.

[Huang and Schroeder, 2006] Huang, B. and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol*, 6:19.

[Porter et al., 2004] Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32(Database issue):D129–33.