A Hybrid Clustering of Protein Binding Sites

Gábor Iván[1,2] Zoltán Szabadka[1,2]
Vince Grolmusz[1,2]

[1]Protein Information Technology Group, Department of Computer Science,
Eötvös University Pázmány Péter stny. 1/C, H-1117 Budapest, Hungary
E-mail: {hugeaux, sinus, grolmusz}@cs.elte.hu
[2]Uratim Ltd. H-4400 Nyíregyháza, Hungary.

January 12, 2010

Abstract

The Protein Data Bank (PDB) contains the description of approximately 27,000 protein-ligand binding sites. Most of the ligands at these sites are biologically active small molecules, affecting the biological function of the protein. Classifying their binding sites may lead to relevant results in drug discovery and design.

Clusters of similar binding sites are created here by a hybrid, sequence and spatial-structure based approach, using the OPTICS clustering algorithm. We defined a dissimilarity-measure: a distance-function on the amino acid sequences of the binding sites. We clustered all the binding sites in PDB according to this distance function, and found that the clusters well characterize the EC codes of those entries that have one.

The color-coded results, containing 20,967 binding sites clustered, are available as html files in three parts at `http://pitgroup.org/seqclust/`.

1 Introduction

In the past few years exploration of the human genome gained the widest publicity. Although somewhat less emphasized, another plenteous bioinformatical resource is the exponentially growing, publicly available Protein Data Bank (PDB) [1], containing more than 55,000 biological structures today.

Three-dimensional structure of smaller molecules – e.g., drug molecules – can usually be calculated from their chemical composition. Several databases exist that contain millions of ligands - an example of this is the freely available ZINC database ([2]) created from catalogues of compound manufacturers.

Contrary to ligands, three dimensional structure of proteins cannot easily be calculated, so the rapid growth of the PDB cannot be overestimated.

Most of the antimicrobial drug molecules act as enzyme inhibitors. Inhibitors need to bind stronger to the enzyme than the substrate of the enzyme; consequently, the chemical and geometrical properties of the binding sites are of utmost importance in drug search and design.

1.1 Our goals

The PDB contains the three-dimensional structures of more than 55,000 entries. In a separate work ([3]), we collected, verified and cleaned the list of approximately 27,000 binding sites, found in the PDB. In the process of identifying these binding sites, we filtered out crystallization artifacts, covalently-bound small molecules, and also took into account broken peptide-chains, modified amino acids, incorrectly labeled HET groups. The resulting cleaned, strictly structured RS-PDB database ([3]) can serve as an input for different data mining algorithms. One such technique of classification is *clustering*. By clustering of binding sites it is possible to create binding site similarity classes. These classes can be useful for classifying protein-ligand interaction.

In this paper we present a fast, sequence-based method for binding site clustering that takes into account amino acid sequences in the close neighborhood of binding sites. Our method is a hybrid in the sense that it uses the sequence information together with the steric data from the PDB in a clearly structured way.

1.2 Previous work

There is a very rich literature describing identification techniques for biological functions from structural protein information by applying highly non-trivial mathematical tools [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Some of these tools were applied for finding or analysing protein-protein interaction network topology [23, 26, 27, 28, 29, 30, 31] or binding sites [23, 32, 33]. A considerable amount of work was also done for devising polypeptide sequence-order independent structural properties [34, 35, 6, 36, 37, 14, 15, 17, 17, 38]. Unlike other binding site clustering solutions in the literature ([39, 40, 41, 42]), we use a hybrid of order-independent and an order-analyzing methods; one of its main features is that it is capable of handling multiple polypeptide chains in the same binding site.

2 Methods

2.1 Binding Site Representation

As a first step, an exact definition of a *binding site* has to be provided. For easy algorithmic handling we store the binding sites found in the PDB in a compact data structure.

The Definition of Binding Sites

A binding sites is defined as a set of atom-pairs; the first atom of the pair belongs to the protein, and the second atom to the bound ligand, such that their distance is equal to the sum of Van der Waals radii, calculated differently for different atom-types. That is, only the pairs within non-covalent binding distance are included in the list. Binding sites, containing covalently bound ligands, are not considered in this work, since the main motivation of ours is to review pharmacologically significant binding sites.

A *binding amino acid (or residue)* is an amino acid with at least one of its atoms in binding atom-pair. A *binding amino acid sequence* is an amino acid sequence that contains at least one *binding amino acid*. Basically, binding sites are represented by storing all the *binding amino acid sequences* of all the protein chains that are present at the particular binding site.

Binding sites were extracted from the RS-PDB database described in ([43] and [3]). By using this definition for binding sites, all amino acids from a given amino acid sequence that have at least one atom contained in an atom pair-set (describing some binding site) can be identified.

Residue Sequence Representation

On *amino acid sequence* we mean sequences consisting of amino acids connected by peptide bonds that are of maximal length (i.e., they cannot be continued with further amino acids on either end).

We note that multiple amino acid sequences might occur in the immediate vicinity of a single binding site, which makes binding site distance/similarity determination fairly complicated. An example of a binding site with four neighboring polypeptide chains can be seen on Figure 1.

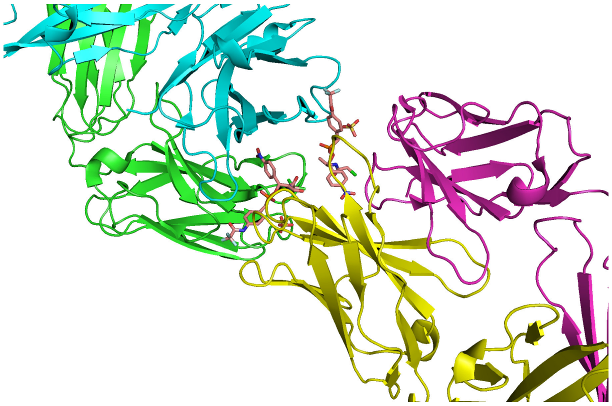

Figure 1: A binding site with four protein chains (PDBID: 1CT8). Each chain is colored diferently.

Binding amino acid sequences were first extracted from the binding sites of the RS-PDB database [43], [3] then they were simplified as follows:

A string was assigned to each amino acid sequence in a binding site. In this string, residues participating in the bond were indicated by their one-character code; non-binding amino acids were indicated by '-' characters. As our purpose was to deal with only the binding sections, the pre- and postfixes consisting of purely non-binding amino acids (or, in our notation, '-' characters), were deleted. Hence all the strings constructed this way start and end with a binding amino acid.

Example. A *binding amino acid sequence* constructed and transformed the way described above (from PDB entry 2BZ6) is shown below:

$$\begin{array}{l}
\quad H - \\
- - - - - - - - - - - TT - - D - \\
- \\
- P - - - - - - - - - - - - - - - \\
- DSCK - - S - - - - - - - - - - - - - - - - - VSWGQGC - - - - - - G
\end{array}$$

2.2 Distance Function

For using a clustering algorithm, we need to define a distance function. The binding sites are represented by all amino acid sequences that participate in the bond with the ligand. Consequently, one have to define the distance of the sequence-sets, situated in the binding sites. This is accomplished first by defining the distance of two sequences (described in section 2.1), then by defining the distance of sequence-sets. The reason for this complexity is the fact that more than one binding sequence can be present in a binding site (cf., Figure 1).

Sequence Comparison Algorithm

For measuring the distances of binding sections of amino acid sequences constructed the way described above, we used a modified version of the algorithm used for calculating *Levenshtein-distance* (furthermore denoted as L). The modifications involved assigning different costs to gaps depending on where they are inserted, while amino acid mismatches were simply penalized by the value 1.

The costs of aligned binding and non-binding amino acids were the following:

- The cost of two aligned, different amino acids is 1.
- The cost of aligned, matching amino acids is zero.

Gaps were penalized as follows:

- Insertion of a gap with a length of one unit (one amino acid) costs gp (*gap penalty*), if the gap is aligned with a non-binding amino acid in the other sequence. If a gap is aligned with a binding amino acid, its cost is 1.
- Insertion of gaps at the end of sequences is only penalized if they are aligned with binding amino acids. Gaps inserted at either end of a sequence have a zero cost, if they are aligned with non-binding amino acids.

It can be shown that the Levenshtein-distance (and also our modified version of it) fulfills the required properties for being a metric. Non-negativity and symmetry can be directly seen from the definition (assuming non-negative costs). It is also obvious that a zero distance can only be achieved by comparing the same objects: $L(x, y) = 0 \iff x = y$ (assuming that every compared sequence starts and ends with a binding amino acid). What is left to prove is the triangle inequality: for every s, t, r strings (binding amino acid sequences),

$$L(s, t) \leq L(s, r) + L(r, t)$$

In other words, the triangle inequality asserts that changing s to t ”via” r cannot cost less than changing s into t directly. As the Levenshtein-distance (by definition) is the minimum possible total cost of operations transforming s into t, and the sequence of operations that transform s into r, and then r into t is also an allowed sequence of operations, it cannot have a lower total cost than $L(s,t)$, as this would contradict to the optimality of $L(s,t)$. (What we may want to prove at this point is that the algorithm we use indeed calculates the defined distance – L.) This reasoning is also applicable to our modified version of the Levenshtein-distance; the only difference is that we have a somewhat more sophisticated set of costs for inserting, deleting and changing characters. We assume that the costs are non-negative, and any binding amino acid sequence compared with our distance function starts and ends with a binding amino acid. We can now reformulate the above defined costs to be used with ”insert”, ”delete”, ”change” operations.

Costs for insertion:

- Inserting a ’-’ character to the end of the sequence: 0.
- Inserting a ’-’ character between the first and last binding amino acid of the sequence: GP.
- Inserting the one-letter code of a binding amino acid: 1.

Costs for deletion:

- Deleting a ’-’ character from the end of the sequence: 0.
- Deleting a ’-’ character between the first and last unchanged binding amino acid of the sequence: GP.
- Deleting the one-letter code of a binding amino acid: 1.

Costs for character changing:

- For matching characters, 0.
- For non-matching characters, 1.

If we want to transform a binding amino acid sequence s into t using the above operations, we cannot expect to get a lower total cost by first transforming s to an arbitrary r and then r to t (compared to directly transforming s to t). This means that the triangle inequality holds.

Binding Site Comparison Method

The input of the distance function described above are two strings that represent amino acid sequences extracted from binding sites. However, our aim is to measure the distance of binding sites, not just single amino acid sequences. We have seen in Section 2.1 on Figure 1 that multiple amino acid sequences might occur in the immediate vicinity of a binding site. Therefore, we also have to define the distance of sequence-sets, representing binding sites.

For this purpose, a complete bipartite graph is defined: This is a graph where the set of vertices can be divided into two disjoint sets A and B such that no edge has both of its endpoints in the same set, while $|A|$ = $|B|$ and the number of edges is always $|A| \cdot |B|$.

- Points of the vertex sets A and B correspond to the amino acid sequences of the first and the second binding sites, respectively. If the numbers of the amino acid sequences are not equal in the two binding sites, amino acid sequences with zero length are added to the smaller set.
- Weights are assigned to all edges of this graph that correspond to the distance of the two amino acid sequences the edge connects. On *distance* we mean the distance defined in Section 2.2.

The distance of the sequence sets A and B is then defined as the minimum weight perfect matching ([44]) in the graph defined above.

We note that by the definition of Section 2.2, the distance of an arbitrary residue-sequence A to a zero-length sequence B is the binding amino acid count of sequence A.

Binding Site Distance Normalization

The expected distance of two randomly generated binding sites will be proportional to the sum of binding amino acids occurring at the binding sites. The maximum achievable distance is always less than the sum of binding amino acids.

The distance of two binding sites calculated using the function described in Section 2.2 does not describe binding site dissimilarity alone. If the distance of two binding sites is 3, it may occur that they have 3 binding amino acids each, hence they can be completely different. On the other hand, a distance of 3 between two binding sites with 30 binding residues each is approximately a ten percent difference, so these binding sites might be almost the same.

Therefore, it is necessary to „normalize" the distances – we did this by dividing all distances by the sum of binding amino acids of the two binding sites in comparison. The result of this operation yields a value between 0 and 1 that can also be interpreted as a percentage of the absolute maximum possible distance of the two binding sites.

2.3 Clustering Algorithm

For data clustering we intended to use an algorithm that is not biased towards even sized and regular shaped clusters.

One algorithm with this properties is DBSCAN ([45]), which is a density-based algorithm. The density of objects is defined with a radius-like ϵ parameter and an object-count lower limit ($minpts$): a neighborhood of some object o is considered dense if there exist at least $minpts$ objects within a less-than-ϵ distance. So, $minpts$ and ϵ are input parameters of the algorithm.

Unfortunately, the clustering structure of many real-data sets cannot be characterized by global density parameters, as quite different local densities may exist in different areas of the data space. The OPTICS (*Ordering Points To Identify the Clustering Structure*) ([46]) algorithm overcomes these difficulties by *ordering* the objects contained in the database, creating the so-called *reachability plot*. The reachability plot is a very clever visualization of high-dimensional clusters. It is basically generated by assigning a value called *reachability distance* to all the objects of the database, while going through the database points in a specific order. The reachability distance is given on the y axes, while the objects (i.e., binding site-representations) are numbered on axes x. Clusters are corresponded to concave regions on the plot. After the creation of the reachability plot, cluster membership assignments can be – among others – created by cutting the reachability plot with a horizontal line furthermore referred to as *cut-level*.

The reachability plot of a small database consisting of binding sites that contain NAD as the ligand can be seen on Figure 2.

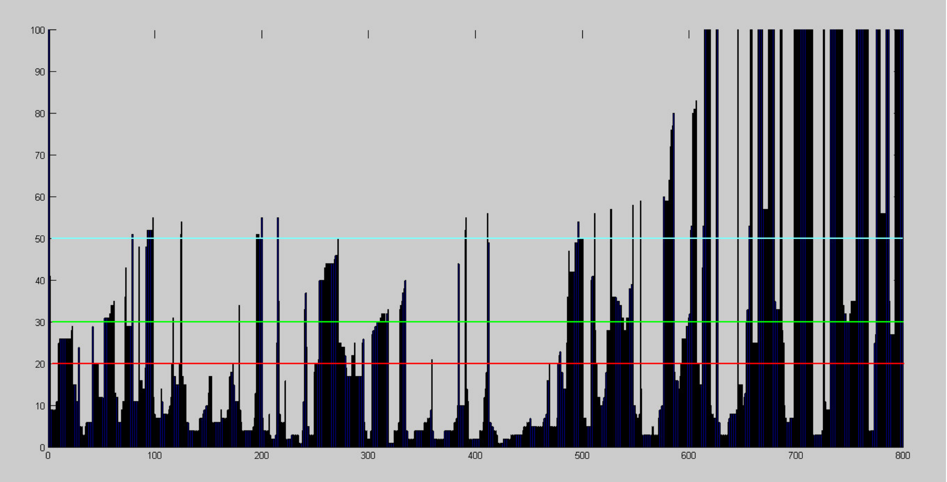

Figure 2: OPTICS reachability plot of a database consisting of 800 binding sites

2.4 Clustering Quality Measurement

Quality of a given clustering depends on several parameters. These include parameters of the distance function used for determining similarity or distance of objects and parameters of the clustering algorithm. In order to get appropriate feedback about the quality of a clustering with a given parameter setting, a quality metrics has to be defined. For this purpose we used the *silhouette coefficient* ([47]). The advantage of the *silhouette coefficient* is that it is completely independent from the type of data being clustered; upon its determination it only uses object distances and cluster membership assignments.

silhouette coefficient	**Clustering quality**
0.00-0.25	Clusters cannot be adequately identified, cluster borders are not obvious
0.25-0.50	Clusters can be identified, but there exist a lot of unclassifiable points (*noise*)
0.50-0.70	Most of the data/points can be classified
0.70-1.00	Excellent distinguishable clusters

Table 1: Cluster quality descriptions based on *silhouette coefficient*'s values by ([47])

Data contained in Table 1 is based on empirical measurements – *silhouette coefficient* values depend greatly on the applied distance function. Therefore, it is questionable to classify clusterings into rigid quality categories based on the *silhouette coefficient* value. However, it is undoubtedly useful for comparing quality of clusterings.

Silhouette coefficient requires the clustering algorithm to assign each binding site to a cluster by definition. Thus, the *silhouette coefficient* value also shows the amount of noise the database contains. The OPTICS algorithm, however, also allows marking some binding sites as ”noise” (thus not putting them into any cluster). It does not seem to be reasonable for binding sites that are ”noise” to be taken into account twice (once, as the OPTICS algorithm marks them, and once at the calculation of *silhouette coefficient*). Therefore, binding sites marked as noise were not taken into account when calculating *silhouette coefficient*. Nevertheless, for the sake of completeness, we will show (Figure 5) how the value of *silhouette coefficient* would change if binding sites marked as noise would be taken into consideration with a silhouette=0 value.

2.5 Database parameters and further settings used in OPTICS algorithm

The OPTICS algorithm was run on a database consisting of 20,967 binding sites. Indistinguishable binding sites – that were assigned exactly to the same *binding amino acid sequence*-sets and ligand identifiers – were contained only once. (The original database – without this kind of redundancy filtering – consisted of 27208 binding sites.) Distance of binding sites was measured with the function described in Section 2.2, using costs introduced in Section 2.2.

3 Results

Our main result is the OPTICS-based clustering of the 20,967 binding sites found. In order to verify the capabilities of the clustering method we need to compare the clusters found with verified biological functions.

3.1 Verification of Results: Biological Relevance

Optimally, in the same cluster proteins of same or closely related functions ought to be assigned. We considered the EC classification of the enzymes, and color-coded the EC numbers in the way that closely related functions got close colors, as given in http://pitgroup.org/seqclust/bsites_AAcodes/EC_colour.html.

The color-coded clusters, together with the ordinal number of the binding site, the PDB ID, the cluster ID and the EC number can be found in three large html files, (Page1, Page2, Page3) under http://pitgroup.org/seqclust/. The clusters correspond to concave regions in the figure.

The deviations of the EC numbers in all the clusters were also computed, and are given in the on-line table http://pitgroup.org/seqclust/bsites_AAcodes/EC_deviation.txt.

We believe that the validation of enzymatic functions through EC numbers shows that the clustering method of ours is an adequate solution for binding-site clustering and classification.

Parameter settings and examples

The parameters used for clustering were the following:

Parameter	**Value**
OPTICS $minpts$	2
OPTICS cut-level	20%
Gap penalty (gp)	$\frac{1}{10}$

We present here as examples four binding sites from the biggest cluster (element count: 448) – see Figure 3. All four proteins are blood clotting factors. The whole cluster is given in the on-line Figure http://pitgroup.org/seqclust/bsites_AAcodes/bsites_optics_M02_No001.html. Note that the whole cluster is colored to blue there, and all the members of the cluster (between line numbers 702 and 1149, cluster ID: 28) have EC numbers of the form 3.4.21.X (serine proteases).

From the second biggest cluster (element count: 188) three binding sites were visualized on Figure 4. The whole cluster is given in the on-line Figure http://pitgroup.org/ seqclust/bsites_optics_M02_No001.html. Note that the whole cluster is colored to deep violet, and almost all the members of the cluster (between line numbers 1224 and 1411) have EC numbers 3.4.23.16 (HIV-1 retropepsins). More analysis on the homogenity of the clusters is given on http://pitgroup.org/seqclust/bsites_AAcodes/EC_deviation.txt.

3.2 Effects of Parameters on Clustering Quality and Cluster Size Distribution

Within our binding site model, distance function and clustering algorithm, three main parameters affected the properties of clustering: OPTICS $minpts$, OPTICS cut-level and distance function's gp. We examined how these parameters affect the quality of clustering measured by *silhouette coefficient*.

- *Effect of parameter gp*: Increasing the gap penalty *gp* slightly improves the quality of the clustering. This is understandable, if we consider that the introduction of a less strict gap penalty function automatically decreases the average distance between clusters.

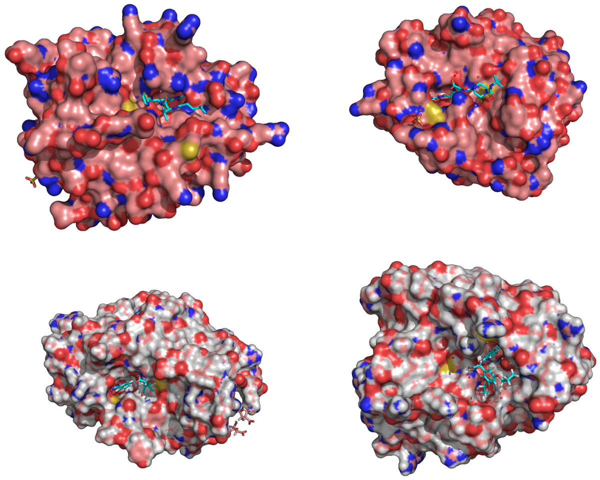

Figure 3: Four binding sites (PDB ID's: 1ZPB, 1RXP, 1C5Z, 2BZ6) from the same cluster. The whole cluster is given in the on-line Figure http://pitgroup.org/seqclust/bsites_optics_M02_No001.html. Note that the whole cluster is colored to blue, and all the members of the cluster (between line numbers 702 and 1149, cluster ID: 28) have EC numbers of the form 3.4.21.X (serine proteases). More analysis on the homogenity of the clusters is given on http://pitgroup.org/seqclust/bsites_AAcodes/EC_deviation.txt

- *Effect of parameter $minpts$*: Increasing $minpts$, two main effects can be observed. On the one hand, an increased $minpts$ means better quality clustering. On the other hand, it also means drastically more binding sites classified as *noise*. The main cause of the latter effect is that the clusters that exist in the database but consist of less than $minpts$ points are not recognized, they are marked as *noise*. Based on this observation, it can be stated that our binding site database contains a lot of small clusters.
- *Effect of parameter OPTICS cut-level*: Increasing cut-level decreases clustering quality, and also the number of binding sites marked as 'noise'. Application of an extremely high cut-level puts almost all binding sites into the same cluster; the quality of such clustering can by no means considered high.

As a conclusion, we state that low minpts and cut-levels yield the best clustering quality (while covering 70-80$\%$ of the binding sites found in PDB).

4 Conclusions

In this paper we presented a fast, sequence-based method capable of classifying the binding sites contained in the publicly available Protein Data Bank. We determined parameter settings yielding a classification with the best quality (measured by *silhouette*

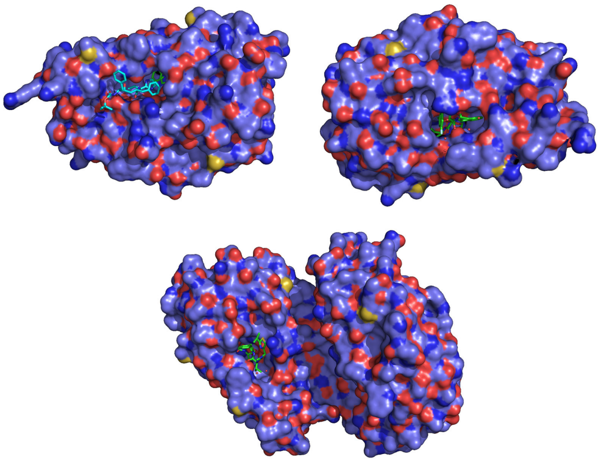

Figure 4: Three binding sites from the same cluster. One site from PDB ID of 1BDL, and two sites on 1W5V, these are HIV-1 proteases. The whole cluster is given in the on-line Figure http://pitgroup.org/seqclust/bsites_AAcodes/bsites_optics_M02_No001.html. Note that the whole cluster is colored to deep violet, and almost all the members of the cluster (between line numbers 1210 and 1435) have EC numbers of the form 3.4.23.16 (HIV-1 retropepsins). More analysis on the homogenity of the clusters is given on http://pitgroup.org/seqclust/bsites_AAcodes/EC_deviation.txt

coefficient). Our main result is a sequence-based approach, derived from 3D structures, used for binding site clustering (rather than three-dimensional binding site structure), that allows *multiple sequences* to occur at each binding site. We also evaluated our clustering results with a large, colored diagram (given at the URL http://www.pitgroup.org/seqclust), where the colors correspond to the EC numbers of the proteins, containing the binding sites. As witnessed by the colored diagram, and also by the numerical deviations given in http://pitgroup.org/seqclust/bsites_AAcodes/EC_deviation.txt, our method has a clear-cut biological significance. The method presented in this work can help reveal evolutionary related binding sites and can also be used for filtering the redundancies (i.e., multiple occurring binding sites) from the PDB. A possible step for further research can be the creation of aggregate sequence-set-profiles for each binding site cluster, generating binding site families similar to the Protein Families Database (*Pfam*) [48, 49].

Acknowledgement. This work was supported by Hungarian Scientific Research Fund (NK-67867), and by the Hungarian National Office for Research and Technology (OMFB-01295/2006 and TB-INTER).

Color	Cut-level
Red	20%
Green	30%
Blue	40%
Cyan	50%
Magenta	60%
Yellow	70%

Table 2: Colours assigned to different OPTICS cut-levels

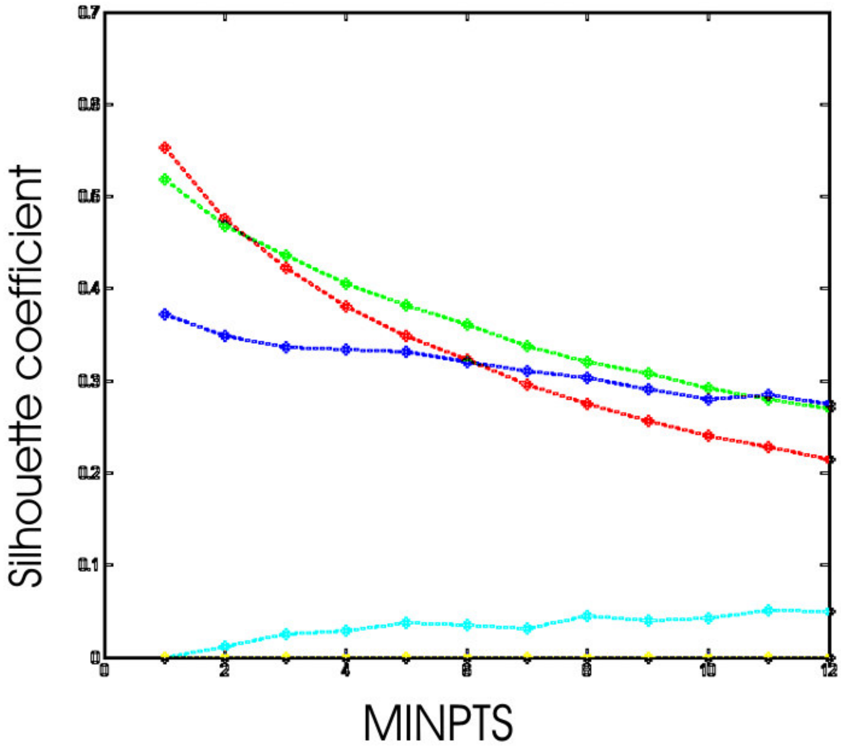

Figure 5: *Silhouette coefficient* dependence on parameter $minpts$, if unclustered binding sites are also taken into account at *silhouette coefficient* determination ($gp = \frac{1}{10}$). The color coding is given in Table 2.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*,

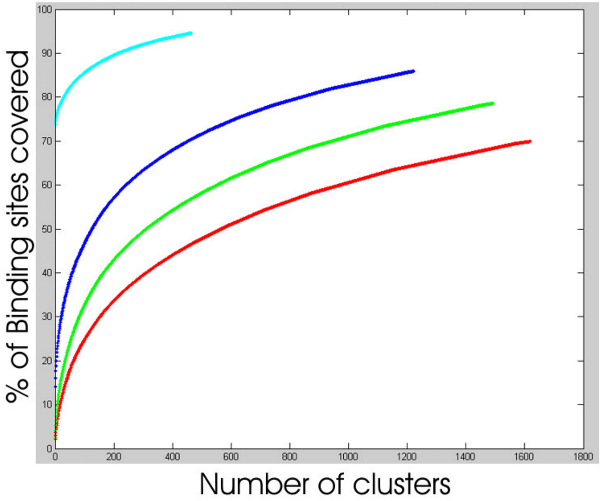

Figure 6: Number of binding sites contained in clusters depending on the number of clusters allowed to be used ($gp = \frac{1}{10}$). The color coding is given in Table 2.

28:235–242, 2000.

[2] John J. Irwin and Brian K. Shoichet. A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Comput. Sci.*, 2005.

[3] Zoltan Szabadka and Vince Grolmusz. Building a structured PDB: The RS-PDB database. In *Proceedings of the 28th IEEE EMBS Annual International Conference, New York, NY, Aug. 30-Sept 3, 2006*, pages 5755–5758, 2006. URL `http://www.cs.elte.hu/%7Egrolmusz/papers/pdb-4.pdf`.

[4] Nelly Andrusier, Ruth Nussinov, and Haim J Wolfson. Firedock: fast interaction refinement in molecular docking. *Proteins*, 69(1):139–159, Oct 2007. URL `http://dx.doi.org/10.1002/prot.21495`.

[5] I.I. Artamonova, G. Frishman, M.S. Gelfand, and D. Frishman. Mining sequence annotation databanks for association patterns. *Bioinformatics*, 21:iii49–iii57, 2005.

[6] E. Azarya-Sprinzak, D. Naor, H. J. Wolfson, and R. Nussinov. Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Eng*, 10(10):1109–1122, Oct 1997.

[7] Hadar Benyamini, Kannan Gunasekaran, Haim Wolfson, and Ruth Nussinov. Fibril modelling by sequence and structure conservation analysis combined with protein docking techniques: beta(2)-microglobulin amyloidosis. *Biochim Biophys*

Acta, 1753(1):121–130, Nov 2005. URL http://dx.doi.org/10.1016/j.bbapap.2005.07.012.

[8] O. Dror, H. Benyamini, R. Nussinov, and H. Wolfson. Mass: multiple structural alignment by secondary structures. *Bioinformatics*, 19 Suppl 1:i95–104, 2003.

[9] Oranit Dror, Dina Schneidman-Duhovny, Alexandra Shulman-Peleg, Ruth Nussinov, Haim J Wolfson, and Roded Sharan. Structural similarity of genetically interacting proteins. *BMC Syst Biol*, 2:69, 2008. URL http://dx.doi.org/10.1186/1752-0509-2-69.

[10] Ugur Emekli, Dina Schneidman-Duhovny, Haim J Wolfson, Ruth Nussinov, and Turkan Haliloglu. Hingeprot: automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227, Mar 2008. URL http://dx.doi.org/10.1002/prot.21613.

[11] Asli Ertekin, Ruth Nussinov, and Turkan Haliloglu. Association of putative concave protein-binding sites with the fluctuation behavior of residues. *Protein Sci*, 15(10):2265–2277, Oct 2006. URL http://dx.doi.org/10.1110/ps.051815006.

[12] D. Fischer, O. Bachar, R. Nussinov, and H. Wolfson. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J Biomol Struct Dyn*, 9(4):769–789, Feb 1992.

[13] D. Fischer, S. L. Lin, H. L. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. *J Mol Biol*, 248(2):459–477, Apr 1995.

[14] D. Fischer, C. J. Tsai, R. Nussinov, and H. Wolfson. A 3d sequence-independent representation of the protein data bank. *Protein Eng*, 8(10):981–997, Oct 1995.

[15] D. Fischer, H. Wolfson, S. L. Lin, and R. Nussinov. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci*, 3(5):769–778, May 1994.

[16] D. Fischer, R. Norel, H. Wolfson, and R. Nussinov. Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition. *Proteins*, 16(3):278–292, Jul 1993. URL http://dx.doi.org/10.1002/prot.340160306.

[17] D. Fischer, H. Wolfson, and R. Nussinov. Spatial, sequence-order-independent structural comparison of alpha/beta proteins: evolutionary implications. *J Biomol Struct Dyn*, 11(2):367–380, Oct 1993.

[18] Alejandra Flores-Ortega, Jordi Casanovas, David Zanuy, Ruth Nussinov, and Carlos Aleman. Conformations of proline analogues having double bonds in the ring. *J Phys Chem B*, 111(19):5475–5482, May 2007. URL http://dx.doi.org/10.1021/jp0712001.

[19] K. Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–443, Nov 2004. URL http://dx.doi.org/10.1002/prot.20232.

[20] K. Gunasekaran, Buyong Ma, and Ruth Nussinov. Triggering loops and enzyme function: identification of loops that trigger and modulate movements. *J Mol Biol*, 332(1):143–159, Sep 2003.

[21] K. Gunasekaran, Buyong Ma, B. Ramakrishnan, Pradman K Qasba, and Ruth Nussinov. Interdependence of backbone flexibility, residue conservation, and enzyme function: a case study on beta1,4-galactosyltransferase-i. *Biochemistry*, 42(13):3674–3687, Apr 2003. URL http://dx.doi.org/10.1021/bi034046r.

[22] Kannan Gunasekaran, Chung-Jung Tsai, and Ruth Nussinov. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol*, 341(5):1327–1341, Aug 2004. URL http://dx.doi.org/10.1016/j.jmb.2004.07.002.

[23] Inbal Halperin, Haim Wolfson, and Ruth Nussinov. Sitelight: binding-site prediction using phage display libraries. *Protein Sci*, 12(7):1344–1359, Jul 2003.

[24] Keren Lasker, Oranit Dror, Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. Ematch: discovery of high resolution structural homologues of protein domains in intermediate resolution cryo-em maps. *IEEE/ACM Trans Comput Biol Bioinform*, 4(1):28–39, 2007. URL http://dx.doi.org/10.1109/TCBB.2007.1003.

[25] Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J Wolfson. Recognition of functional sites in protein structures. *J Mol Biol*, 339(3):607–633, Jun 2004. URL http://dx.doi.org/10.1016/j.jmb.2004.04.012.

[26] Yuval Inbar, Hadar Benyamini, Ruth Nussinov, and Haim J Wolfson. Prediction of multimolecular assemblies by multiple docking. *J Mol Biol*, 349(2):435–447, Jun 2005. URL http://dx.doi.org/10.1016/j.jmb.2005.03.039.

[27] Yuval Inbar, Hadar Benyamini, Ruth Nussinov, and Haim J Wolfson. Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics*, 19 Suppl 1:i158–i168, 2003.

[28] O. Keskin, A. Gursoy, B. Ma, and R. Nussinov. Towards drugs targeting multiple proteins in a systems biology approach. *Curr Top Med Chem*, 7(10):943–951, 2007.

[29] Ozlem Keskin, Ruth Nussinov, and Attila Gursoy. Prism: protein-protein interaction prediction by structural matching. *Methods Mol Biol*, 484:505–521, 2008. URL http://dx.doi.org/10.1007/978-1-59745-398-1_30.

[30] Xiang Li, Ozlem Keskin, Buyong Ma, Ruth Nussinov, and Jie Liang. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J Mol Biol*, 344(3):781–795, Nov 2004. URL http://dx.doi.org/10.1016/j.jmb.2004.09.051.

[31] Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J Wolfson. Siteengines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res*, 33(Web Server issue):W337–W341, Jul 2005. URL http://dx.doi.org/10.1093/nar/gki482.

[32] Ozlem Keskin and Ruth Nussinov. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, 15(3):341–354, Mar 2007. URL http://dx.doi.org/10.1016/j.str.2007.01.007.

[33] Ozlem Keskin and Ruth Nussinov. Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng Des Sel*, 18(1):11–24, Jan 2005. URL http://dx.doi.org/10.1093/protein/gzh095.

[34] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, 260(4):604–620, Jul 1996. URL http://dx.doi.org/10.1006/jmbi.1996.0424.

[35] V. Alesker, R. Nussinov, and H. J. Wolfson. Detection of non-topological motifs in protein structures. *Protein Eng*, 9(12):1103–1119, Dec 1996.

[36] O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Eng*, 6(3):279–288, Apr 1993.

[37] Oranit Dror, Hadar Benyamini, Ruth Nussinov, and Haim J Wolfson. Multiple structural alignment by secondary structures: algorithm and applications. *Protein Sci*, 12(11):2492–2507, Nov 2003.

[38] N. Leibowitz, Z. Y. Fligelman, R. Nussinov, and H. J. Wolfson. Automated multiple structure alignment and detection of a common substructural motif. *Proteins*, 43(3):235–245, May 2001.

[39] Nicola D Gold and Richard M Jackson. Sitesbase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res*, 34(Database issue):D231–D234, Jan 2006. URL http://dx.doi.org/10.1093/nar/gkj062.

[40] Sarah L Kinnings and Richard M Jackson. Binding site similarity analysis for the functional classification of the protein kinase family. *J Chem Inf Model*, 49(2):318–329, Feb 2009.

[41] Daniel Kuhn, Nils Weskamp, Eyke HĂźllermeier, and Gerhard Klebe. Functional classification of protein kinase binding sites using cavbase. *ChemMedChem*, 2 (10):1432–1447, Oct 2007. URL `http://dx.doi.org/10.1002/cmdc.200700075`.

[42] Akira R Kinjo and Haruki Nakamura. Comprehensive structural classification of ligand-binding motifs in proteins. *Structure*, 17(2):234–246, Feb 2009. URL `http://dx.doi.org/10.1016/j.str.2008.11.009`.

[43] Zoltan Szabadka and Vince Grolmusz. High throughput processing of the structural information in the protein data bank. *J Mol Graph Model*, 25(6):831–836, Mar 2007. URL `http://dx.doi.org/10.1016/j.jmgm.2006.08.004`.

[44] L. Lovász and M. D. Plummer. *Matching theory*, volume 121 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 1986. ISBN 0-444-87916-1. xxvii+544 pp. Annals of Discrete Mathematics, 29.

[45] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press,*, 1996.

[46] M. Ankerst, M. M.Breunig, H. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD '99 Int. Conf. on Management of Data, Philadelphia PA*, 1999.

[47] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[48] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322, Jan 1998.

[49] E. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420, Jul 1997.