Pattern Recognition Strategies for Molecular Surfaces: III. Binding Site Prediction with a Neural Network

MATTHIAS KEIL,[1] THOMAS E. EXNER,[1,2] JÜRGEN BRICKMANN[1,3]
[1]*Department of Physical Chemistry, Darmstadt University of Technology, 64287 Darmstadt, Germany*
[2]*Mathematical Chemistry Research Unit, Department of Chemistry, University of Saskatchewan, 110 Science Place, Saskatoon S7N 5C9, Canada*
[3]*Darmstadt Center of Scientific Computing, 64287 Darmstadt, Germany*

Received 22 May 2003; Accepted 10 July 2003

Abstract: An algorithm for the identification of possible binding sites of biomolecules, which are represented as regions of the molecular surface, is introduced. The algorithm is based on the segmentation of the molecular surface into overlapping patches as described in the first article of this series.[1] The properties of these patches (calculated on the basis of physical and chemical properties) are used for the analysis of the molecular surfaces of 7821 proteins and protein complexes. Special attention is drawn to known protein binding sites. A binding site identification algorithm is realized on the basis of the calculated data using a neural network strategy. The neural network is able to classify surface patches as protein–protein, protein–DNA, protein–ligand, or nonbinding sites. To show the capability of the algorithm, results of the surface analysis and the predictions are presented and discussed with representative examples.

 J Comput Chem 25: 779–789, 2004

Key words: molecular surface; molecular recognition; binding sites; neural network

Introduction

In the first article of this series,[1] the importance of molecular recognition in chemistry in general and in biochemical systems in particular has been illustrated and a new method for the description of molecular features based on the concept of molecular surfaces and on fuzzy set theory has been suggested. In the second part,[2] we demonstrated that this new description can be used, at least in a first approximate manner, in the search for the relative orientation of the molecules in the complex, i.e., for the treatment of the so-called geometric docking problem. In this third part, we approach the molecular recognition problem from a different direction. While in the docking problem both partners forming the biomolecular complex are known, we concentrate here on one partner, the protein, only and try to find the regions of the molecular surface involved in the molecular recognition process without any information of the second molecule. This is even more complicated to handle then for the case where one is dealing with the complementarity problem of two partners. Because only one binding partner is known the identification of the binding sites cannot rely on specific interactions, like shape complementarity, hydrogen bonds, or salt bridges, between the molecules. A method has to be developed that can identify slight differences of the molecular properties at the binding site in relation to other regions of the molecular surface.

In the last years, the structure of thousands of proteins and protein complexes were solved experimentally and stored in the public Protein Data Bank (PDB).[3,4] Several publications use the 3D data provided in this database to analyze the properties of proteins in general and of binding sites in particular.[5–23] These analyses include:

- Amino acid composition of different protein regions.[7,8,10,12–17]
- Hydrophobicity of the protein surface, binding site, and core region.[7,8,10,12–17,20–22]
- Electrostatic complementarity of binding regions.[18–20,22]
- Intermolecular hydrogen binding[7,8,10,16,19,20] and salt bridges[18–20] in protein complexes.
- Size, shape, and sterical complementarity of binding sites.[7,8,10–16,19,20,22,23]

All these investigations stress the large complementarity of the shape as well as the complementarity of the molecular properties between the two molecular partners building a complex. They also

Correspondence to: J. Brickmann; e-mail: brick@pc.chemie.th-darmstadt.de

M. Keil's present address is Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144.

show that binding sites share some common features but depend on the molecule to which the protein is bound. In this way, protein–protein interfaces are more hydrophobic than other parts of the protein surface, even if they are less hydrophobic than the core region.[6,8,10,15] Thus, on the one hand, hydrophobic interactions are important not only in folding resulting in the tertiary structure but also in complex formation resulting in the quaternary structure. On the other hand, the number of charged amino acids is significantly increased in binding sites compared to the core region. Elcock et al.[5] related this to the fact that, while charged groups destabilize the hydrophobic core of a protein, these groups can stabilize complexes by building salt bridges across the binding interface. To bind to a DNA molecule, the binding site of a protein has to be positive polarized to favor interactions with the negative charges of the phosphate backbone.[24] Thus, positive charged amino acids, like arginine, occur more often in DNA binding sites. In addition, the number of hydrogen bond donors is also increased. Binding sites for small molecules are mainly characterized by their concave shape. They are located in deep cavities, like bags and clefts of the molecular surface.[25–31]

These special features open the possibility to find binding sites by comparing the molecular properties at different parts of the protein surfaces and find those that differ from the average. Because of the great importance of enzyme–substrate and enzyme–inhibitor complexes in the rational drug design and the significant shape of their complex interfaces, most theoretical investigations to identify binding sites were carried out with this class of complexes.[25–36] In the algorithm of Peters et al.,[28] the protein shape is described by the alpha-shape algorithm of Edelsbrunner and Mücke,[37] which generates a whole family of shapes with a different level of complexity. By comparing two shapes of different resolutions, cavities on the surface of the protein, which are possibly involved in ligand binding, are found. The PASS algorithm by Brady and Stouten[26] fills cavities with a set of spheres and identifies a few of these spheres that most likely represent the centers of binding pockets. This is done by adding multiple layers of probe spheres and retaining only those spheres with a low solvent exposure. Other methods, so-called grid-based algorithms,[33,36] rely on neighborhood information to identify clusters of grid points that are surrounded by grid points lying inside the protein.

The program GRID by Goodford[38] also places the protein into a grid but calculates interaction energies with different probes on the grid points. A number of different probes representing different functional groups including positive charges, water molecules, aromatic carbons, hydrogen bond donors, and acceptors can be chosen in this and many similar methods.[27,33,39–42] Clusters of points with attractive interactions to the protein can then be identified as possible binding regions. Ruppert et al.[31] followed a similar strategy by surrounding the protein with probes indicating possible hydrogen bonds or favorable hydrophobic interactions with ligands. These probes are ranked by a scoring function. The regions containing many high-scoring probes are identified as places where a part of a hypothetical ligand might bind with high affinity.

Many studies of protein–protein complexes discuss the increased hydrophobicity of binding sites of these complexes. Young et al.[21] combined hydrophobic amino acids at the protein surface to clusters. These clusters were used to identify 65% of the binding sites of 38 protein–protein complexes. MacCallum et al.[43] studied the amino acid composition of 26 antibody–antigen complexes and used this information for the detection of antigen binding sites. The most similar approach to ours is the one by Jones et al,[12,13] who used overlapping surface patches and six parameters describing the amino acid composition, the hydrophobicity, and the shape of these patches to characterize the molecular surface in a similar manner to the investigations described in this article. The analysis of these parameters led to a method that was able to distinguish between binding and nonbonding patches with an accuracy of 66% to identify the correct binding site in 59 test cases representing a variety of complex types like oligomers, enzyme–inhibitor, and antibody–antigen complexes.

All methods mentioned so far, except the last, concentrate on one specific class of biomolecular complexes, like protein–protein or protein–ligand complexes. In this article, we introduce a new scheme to identify binding sites suitable for a large variety of different molecules interacting with proteins. To accomplish this, we use the fuzzy fragmentation method described in the first two articles of this series in combination with a neural network for the analysis of the properties of these domains. The article is organized as follows: In the next section, we describe the analysis of the molecular properties of protein surfaces based on all structure information available in the PDB. These results are then used for the training of a neural network (NN), which is described in the third section. The identification rate of the neural network and a method to visualize these results on the molecular surface are presented and discussed in the Results section. The last section provides some concluding remarks and a short preview of possible applications and improvements of the method.

Molecular Properties of Binding Sites

Our investigations are based on the PDB[44] as published in release 90, October 1999. This release was composed out of 10,213 structures. To handle this large amount of structures, an automatic procedure was developed. The CHARMM program package (version 24)[45] was used to add and optimize missing nuclei positions especially of the hydrogen atoms, which cannot be determined by X-ray crystallographic methods for these macromolecules and are therefore not included in most entries of the PDB. The control scripts for CHARMM were generated automatically and the output was converted to a format, which can be used for further investigations. Due to the limitation of the CHARMM force field, only proteins composed of the 20 standard amino acids and DNA molecules can be treated with this program. All database entries with nonstandard amino acids were excluded by the automatic procedure. In addition, entries containing only DNA molecules or carbohydrates are not considered any further. At the end, 7821 structures (76%) successfully pass this limiting part of the automatic procedure.

For the further characterization of the molecules, we used physicochemical as well as topographical properties calculated on or projected onto the molecular surface. These properties are the electrostatic potential, the local lipophilicity, the hydrogen bond donor and hydrogen bond acceptor density, the surface topography

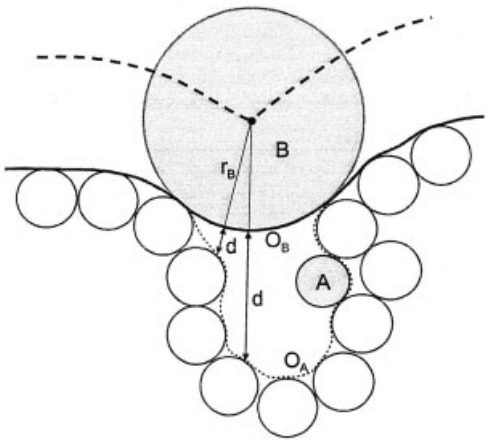

Figure 1. Schematic representation of the calculation of the cavity depth. The two surfaces O_A and O_B are shown as dotted and bold lines with the probe spheres A and B, respectively. The nonfilled spheres represent the atoms of the molecule. Two of the distances d defining the cavity depth are also displayed.

index (STI), and the cavity depth. The concept of a solvent-accessible surface[46–49] and the definitions and calculation methods of all these properties except the cavity depth are well established and are described in the first article of this series.[1] The cavity depth will be described below. The preparation of the surface and calculation of the properties follow a standard procedure.[50–52] First, the molecular surface was created by a grid-based method,[53] resulting in a triangular mesh. Then, the properties are calculated at the positions of the points of the mesh. The atom-centered partial charges and the information on which atoms can form hydrogen bonds are taken from the CHARMM force field[45] and used in the calculations of the properties. For the calculation of the local lipophilicity, the partial lipophilicity values are assigned to each atom according to the definition of Ghose and Crippen[54,55] and Viswanadhan et al.[56] using the MOLFESD program.[57]

The surfaces are then divided into domains with the hydrogen bond density as basic variable following the procedure described in the first article of this series.[1] The minimum and maximum sizes of the domains was set to 50 and 250 Å^2, respectively. This results in a total number of 1,255,853 surface domains for the 7821 structures of the PDB treated by the automatic procedure.

Cavity Depth

In this work, we use two different properties to describe the topographical features of the molecular surface. The first, the STI, quantifies the concavity/convexity of the surface in the area around a surface point as presented in part I.[1] The second is the cavity depth measuring the solvents exposure of the surface, i.e., how deep a surface point is located inside a cavity of the molecule. The calculation of this property is based on the difference of two molecular surfaces (see Fig. 1). In addition to the given molecular surface O_A (generated with the standard radius of the probe sphere of 1.4 Å representing a water molecule) a second molecular surface O_B with a much larger probe sphere radius is generated. In this work a radius of 6.0 Å was used. Because the larger probe sphere cannot penetrate small cavities (cavity entrance smaller than 12 Å in diameter), the surface points of these cavities are inside O_B and the difference between the surfaces O_A and O_B is a measure for the depth of the cavity. The shortest distance of each surface point of O_A to the center of a virtual probe sphere B—used for the generation of O_B— is determined. The surface depth is then defined as this distance minus the probe radius r_B, which corresponds to the distance from O_A to O_B along the line connecting the surface point with the probe sphere center (see Fig. 1). Thus, all surface points of O_A not inside a cavity have a cavity depth near or equal 0 and the value of the cavity depth increases with the extent to which the surface point lies inside a cavity of the molecule.

Assignment of Complex Classes

In the following we compare the molecular properties in different parts of the protein surface to identify possible binding sites by finding differences of the properties between nonbonding and binding sites for various types of complexes. Therefore, each point of the molecular surface is characterized by the ability to build a complex with other molecules. This is done by calculating the minimum distance of each surface point of a protein to the surfaces of all other molecules included in the complex. A point of the protein surface is defined as part of a binding site if another molecule is found in a distance less than the given cutoff of 1.5 Å. To specify the type of complex, the so-called complex class, all complex partners are subdivided into three classes:

1. Peptides with more than eight amino acids and proteins belong to the class protein.
2. DNA and RNA molecules belong to the class DNA.
3. All other molecules are classified as ligands.

If no molecule is found within the cutoff distance from the surface point, it is classified as nonbonding. Following this scheme, each point of a protein surface is classified to belong to one of these four complex classes: protein–protein, protein–DNA, and protein–ligand or nonbonding.

In addition, the complex class of each surface domain is determined by examining the complex classes of all surface points belonging to this domain. If less than 20% of the surface points of one domain belong to a specific surface class, the domain is classified as nonbonding. Otherwise, the complex class with the highest contribution to the surface area is assigned to the domain. This small value of 20% was used to ensure that domains on the border of the binding sites are also included in the characterization of this site. Finally, for each surface domain representative values of the six described molecular properties are determined by calculating their mean values over all points belonging to one surface domain.

Statistical Investigations of Binding Sites

The properties of the molecular surfaces of all protein structures were statistically studied in great detail.[58] Here, we only present

Table 1. Mean Molecular Property Values of the Surface Patches in Different Regions of the Protein Surfaces (Standard Deviations of the Values are Listed as Well).

Property	Complete surface	Protein–protein binding sites	Protein–DNA binding sites	Protein–ligand binding sites
Number of surface domains	1,255,853	147,784	3273	10,192
Size of domains (Å^2)	261.7 ± 54.1	264.7 ± 50.6	266.5 ± 56.4	250.0 ± 61.9
Electrostatic potential (kcal/mol · e)	−1.9 ± 34.8	−0.7 ± 32.1	50.3 ± 33.1	9.0 ± 42.1
Local lipophilicity	−0.0633 ± 0.0383	−0.0424 ± 0.0436	−0.0626 ± 0.0339	−0.0264 ± 0.0488
H-acceptor density (Å^{-2})	0.0229 ± 0.0064	0.0199 ± 0.0062	0.0188 ± 0.0067	0.0244 ± 0.0095
H-donor density (Å^{-2})	0.0212 ± 0.0086	0.0182 ± 0.0084	0.0312 ± 0.0090	0.0251 ± 0.0131
Cavity depth (Å)	1.30 ± 1.25	1.24 ± 0.94	1.76 ± 1.48	5.07 ± 3.21
STI	2.28 ± 0.33	2.33 ± 0.31	2.23 ± 0.32	1.63 ± 0.47

the results, which are important for the development of the method to identify binding sites with a neural network.

Table 1 and Figure 2 show the mean values of the molecular properties of the surface domains. In the diagrams, the mean property values of all surface domains as well as the mean property values in the specific binding sites are presented. Protein–protein binding sites are characterized mainly through their higher lipophilicity as compared to the total surface. This confirms the results, mentioned in the introduction, that hydrophobic interactions are important not only in folding resulting in the tertiary structure but also in complex formation resulting in the quaternary structure of proteins. The hydrogen acceptor and hydrogen donor density is reduced in protein–protein interfaces, which proves that the binding sites are composed for a large part of apolar and therefore lipophilic atoms. In the same sense, the electrostatic potential is almost 0 due to the high electrostatic complementarity of the binding sites but also due to the apolar character of lipophilic regions.

As already mentioned, small ligands bind in deep holes or clefts inside the protein. This is demonstrated by both topographical properties. With a mean cavity depth of more than 5.0 Å, parts of the molecular surfaces building ligand binding sites are almost five times deeper buried inside the protein as the average surface part. Even if the differences in the STIs are not so dramatic, ligand binding sites are concave compared to other parts of the surface, which are on average convex. This also favors the location inside of bags and clefts. The lipophilicity of ligand interfaces is also increased. This is not a necessary feature of the binding sites but is caused by the location near the lipophilic core of the protein. All molecules not proteins, DNA, or RNA are classified as ligands. Therefore, this group is the most diverse and every binding site is optimized for a specific ligand. This can be seen by the large deviation of the lipophilicity as well as the electrostatic potential in different binding sites (see Table 1).

Protein–DNA interfaces are characterized by high positive electrostatic potential. This makes attractive interactions with the negatively charged phosphate groups of the DNA/RNA backbone possible. These specific interactions are also observable in the diagrams showing the hydrogen bond acceptor and hydrogen bond donor densities. While the number of hydrogen bond acceptors is reduced in these interfaces, the number of donors is increased. Thus, the protein can build more hydrogen bonds with the oxygen atoms of the phosphate groups acting as hydrogen bond acceptors.

Neural Network

In this work, the identification of possible binding sites is accomplished using an NN based on the specific properties of binding sites described above. NNs try to mimic the procedure of the human brain by using a large number of simple units, called neurons, and a training algorithm to process data and find nonlinear relationships within this data. Due to the learning capacity of the network, complex pattern recognition problems can be treated for which standard methods cannot be used. This opens many new applications in chemistry[59–61] among many other areas. In the group of Gasteiger,[61,62] e.g., NNs are used to compare geometric and electronic properties of molecules. Sadowski and Kubinyi[63] developed an NN that is able to distinguish between drugs and nondrugs based on the extraction of knowledge from large databases and encoding the molecular structures using atom type descriptors. Promising applications for the prediction of secondary structure elements of proteins based on their amino acid sequence have been reported as well.[64–66]

NN Architecture

The Stuttgart Neural Network Simulator (SNNS, version 4.2)[67] was used to build the NN applied here. We used a simple feed-forward network with one hidden layer, which is shown in Figure 3. The input layer is composed of 17, the hidden layer of 13 and the output layer of 4 units. Each of the output neurons is assigned to one of the previously defined complex classes protein–protein, protein–DNA, protein–ligand, and nonbonding. Each layer is fully connected with the following one and a sigmoid activation function is applied to determine the activation of the next layer from the output of the layer above. For the training the standard backpropagation rule was used.[59,61,68]

Processing of the Molecular Properties

The six molecular properties—electrostatic potential, local lipophilicity, hydrogen bond donor and hydrogen bond acceptor

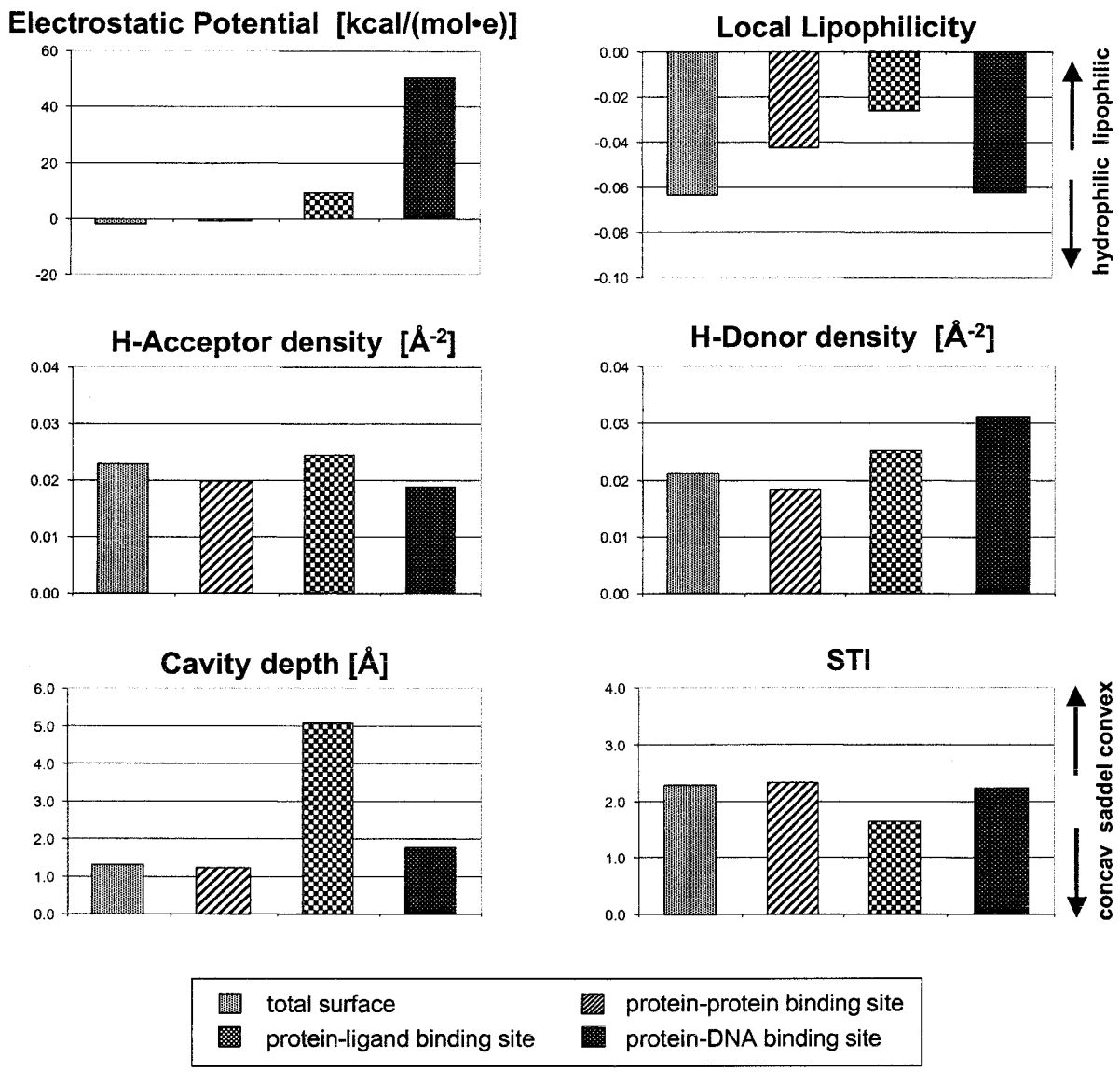

Figure 2. Comparison of the mean molecular properties in the binding regions with the mean property values of the total surface.

density, cavity depth, and STI—are used as input variables. Four of the molecular properties (hydrogen bond donor, hydrogen bond acceptor density, cavity depth, and STI) were represented by the mean values of these properties. The electrostatic potential and the local lipophilicity are represented by two mean values. The first is the mean value of all surface points with positive electrostatic potential and all lipophilic surface points of one domain, respectively. The second mean value is calculated over the electronegative and hydrophilic surface points, respectively.

Each property is processed by up to four input units and the incoming signal of each input unit is determined by piecewise linear functions shown in Figure 4. The numbers of input units assigned to each molecular property as well as the values defining the actual shape of the input signal functions are summarized in Table 2.

Training

For the training of the NN, the complete dataset of 1,255,853 surface domains was subdivided into a training and a test set. In our first trials using random selected domains for the training set, we were confronted with the problem that the NN identifies almost all domains as nonbonding. The reason for this is that almost 90% of the domains of the complete data set are classified as nonbonding. The identification rate of these NNs is therefore almost 90% because all nonbonding domains are identified correctly. But, no

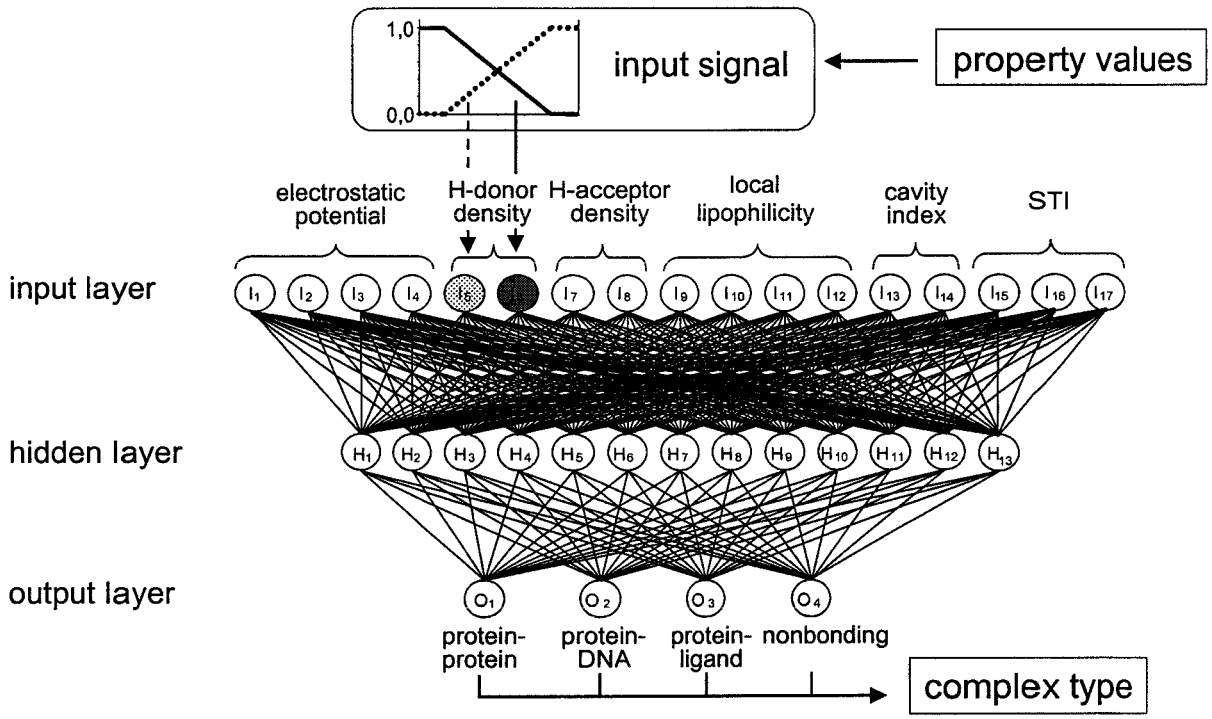

Figure 3. Architecture of the feed-forward neural network with one input, one output, and one hidden layer. The assignment to the molecular properties to the input units and the generation of the input signals are also shown.

binding interface can be predicted and so these NNs are not usable for the detection of binding sites. To circumvent this problem, the percentage of binding domains was increased in the training set by using only the domains of known protein complexes. In addition, the set was limited to selected complexes to reduce the time needed for the training phase. At the end, a training set of 13,994 domains was created in which approximate 43% of the domains are involved in the binding to other molecules. The test set was composed of domains of all other proteins.

The NN was trained until no improvement in the successful identification of binding sites could be obtained. The identification rate of the NN is measured by a function combining the identification rates for each complex class and the overall identification rate for all classes. The identification rate for a specific class is defined as

$$S_\alpha = \frac{m_\alpha}{n_\alpha}, \tag{1}$$

with S_α the identification rate of complex class α (α = protein–protein, protein–DNA, protein–ligand, or nonbonding), m_α the

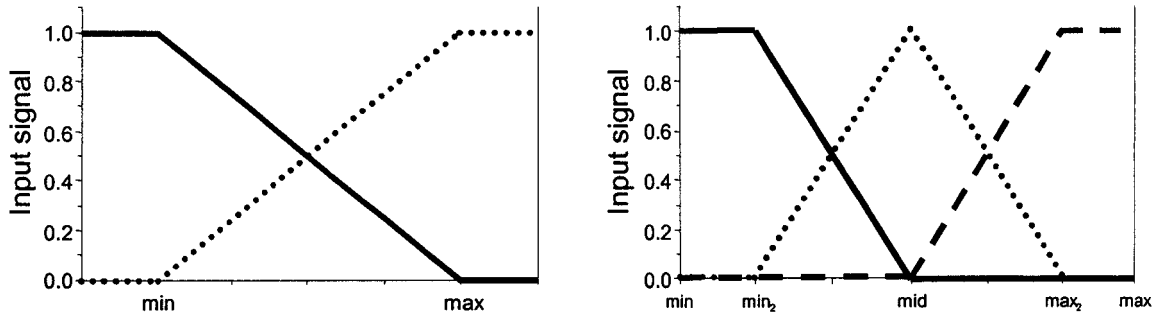

Figure 4. Input signal functions for a molecular property represented by two or three input units respectively. (—), input signal of first unit; (· · ·), input signal of second unit; (- - -), input signal of third unit.

Table 2. Numbers of Input Units and Values Defining the Activation Functions Used in the NN for Each Molecular Property.

Property	No. of input units	Min	Mid	Max
Electrostatic Potentials (positive part)	2	0.0		100.0
Electrostatic Potentials (negative part)	2	−100.0		0.0
Liphophilicity (lipophilic part)	2	0.00		0.10
Liphophilicity (hydrophilic part)	2	−0.25		0.00
Hydrogen bond donor density	2	0.00		0.06
Hydrogen bond acceptor density	2	0.00		0.06
Cavity depth	2	0.0		6.0
STI	3	1.0	2.0	3.0

number of correctly identified domains of complex class α, and n_α the number of domains of complex class α.

In the data set, there are many surface domains for which no strong output activations can be generated by the neural network. But, for most cases the highest output activation, even if it has a small value, corresponds to the correct complex class. Therefore, a domain is considered correctly identified if the activation of the complex class to which the domain was assigned is the highest of all activations of the output neurons and is higher than 0.15. By using this small cutoff value, the identification rates of the NN were increased drastically compared to higher cutoff values above.

The overall identification rate is the sum of the identification rates of the individual complex classes weighted by the numbers of domains assigned to each of these classes:

$$S_{\text{total}} = \frac{m_{\text{protein}} + m_{\text{DNA}} + m_{\text{ligand}} + m_{\text{nonbonding}}}{n_{\text{total}}} = \frac{n_{\text{protein}}}{n_{\text{total}}} \cdot S_{\text{protein}} + \frac{n_{\text{DNA}}}{n_{\text{total}}} \cdot S_{\text{DNA}} + \frac{n_{\text{ligand}}}{n_{\text{total}}} \cdot S_{\text{ligand}} + \frac{n_{\text{nonbonding}}}{n_{\text{total}}} \cdot S_{\text{nonbonding}}, \quad (2)$$

with S_{total} the overall identification rate of NN and n_{total} the number of domains of all complex classes α.

Results and Discussion

Classification of Surface Domains

The identification rates of the NN after its training are shown in Table 3 (first column). The NN is capable of predicting the correct class for 76% of all surface domains in the PDB. The NN detects well the nonbonding parts of the molecular surface (80%) but has problems in distinguishing between protein, DNA, and ligand binding sites. In some way, this is not unexpected and is due to the somewhat arbitrary assignment of the complex classes, e.g., the active sites of enzymes are designed to catalyze the reactions of substrates, which can be small. Therefore, the binding sites of these enzymes are more like the ones of proteins binding small molecules, which we have included in the protein–ligand complex class in this work. But, in the cell enzymatic reactions are often regulated by the inhibition with small proteins. In this case, the same binding site is assigned to the protein–protein class according to our classification scheme.

To circumvent this problem, we decided to take also the second best prediction of the NN into account for the evaluation of the NN. Using this approach a domain is considered correctly identified if the activation of the complex class to which the domain was assigned is the highest or second highest of all activations of the output neurons and is higher than 0.15. The second highest activation is only considered as a valid prediction if the difference to the highest activation value is less than 0.5. This approach enhances the identification rates of the NN significantly (see the second column in Table 3).

The NN identifies the nonbinding surface part the best. The identification of ligand binding sites is also high. This can be explained by the special location of the ligand binding sites. They are mostly located in deep pockets and ridges of the protein surface (see Fig. 2), which clearly differentiate them from the other surface parts.

Visual Binding Site Detection

Because binding sites span a large area of the molecular surface, more than a single surface domain is part of them. Therefore, it is important not only to identify the complex class of specific domains but also to find clusters of overlapping domains on the molecular surface belonging to the same complex class. On the one hand, a binding site is more favorable if the number of domains included in the corresponding cluster increases. In this way, a binding site can be identified even if no domain has a high activation of the neuron corresponding to the complex class but a number of domains have activations high enough that a large cluster can be found. On the other hand, it is not likely that a single domain with a high activity of the output neuron of a specific complex class surrounded by nonbonding domains is able to build strong interactions with another molecule and, thus, no complex can be formed by this domain alone.

One possibility to find clusters is to project the complex class information for each domain back on the molecular surface. In this sense, three new properties corresponding to one of the three binding complex classes are calculated at each point of the molecular surface. These new properties are defined as the output

Table 3. Identification Rates Using Only the Best Prediction and the Two Best Predictions of the NN.

Identification rate based on . . .	S_{total}	S_{protein}	S_{DNA}	S_{ligand}	$S_{\text{nonbonding}}$
Best prediction only	76%	44%	41%	62%	80%
Best and second best predictions	90%	71%	64%	74%	93%

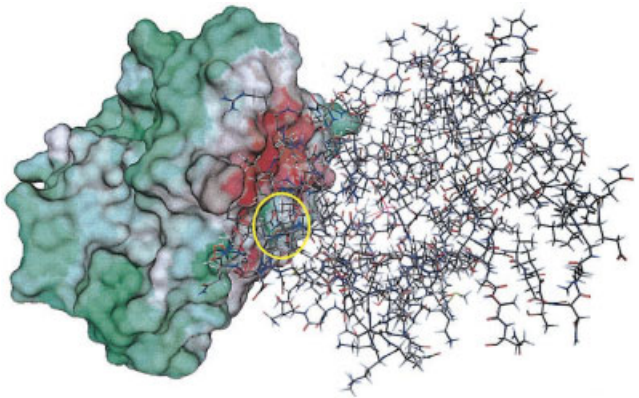

Figure 5. Binding site of the thymidilate kinase dimer. The property describing the ability to form protein–protein complexes is shown color-coded on the molecular surface of the first protein (red, high probability for protein–protein complex formation; green, low probability). The second protein is represented as a stick model. The yellow circle surrounds the region where the NN is not able to identify the binding region correctly.

neuron activations for a particular complex class averaged over all domains to which the surface point has a membership larger than 0, i.e., to which the surface point belongs. By the visual inspection of the properties on the molecular surface, binding sites can be identified as large regions with a high value of these properties.

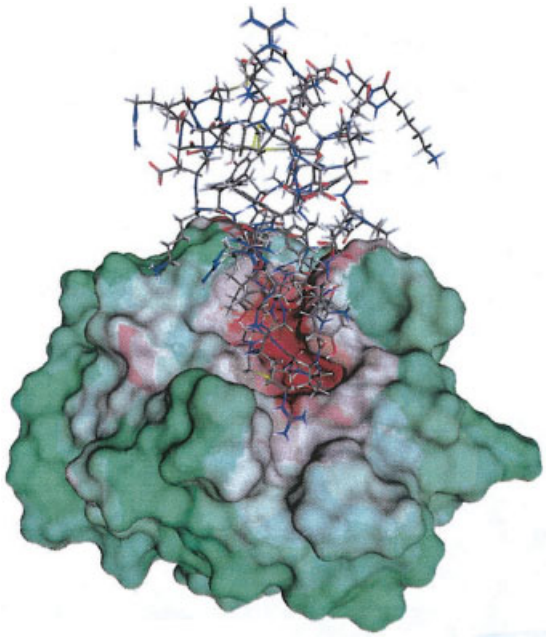

Figure 6. Binding site of the β-trypsin–BPTI complex. The property describing the ability to form protein–protein complexes is shown color-coded on the molecular surface of β-trypsin (red, high probability for complex formation; green, low probability). BPTI is represented as a stick model.

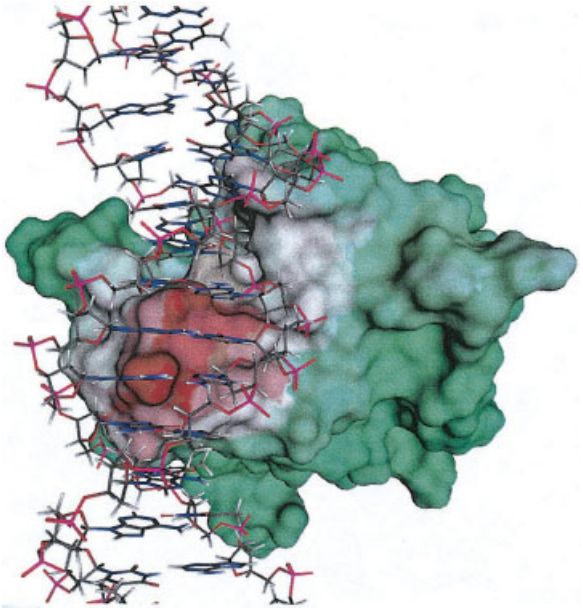

Figure 7. Binding site of the p53 tumor suppressor complex. The property describing the ability to form protein–DNA complexes is shown color-coded on the molecular surface of the chain B of the p53–DNA complex (red, high probability for complex formation; green, low probability). The part of the DNA to which the p53 protein binds is represented as a stick model.

This is shown by four examples in Figures 5–8. None of the example protein complexes was included in the training set of the NN. Thus, these examples demonstrate the ability of the NN to identify unknown binding sites.

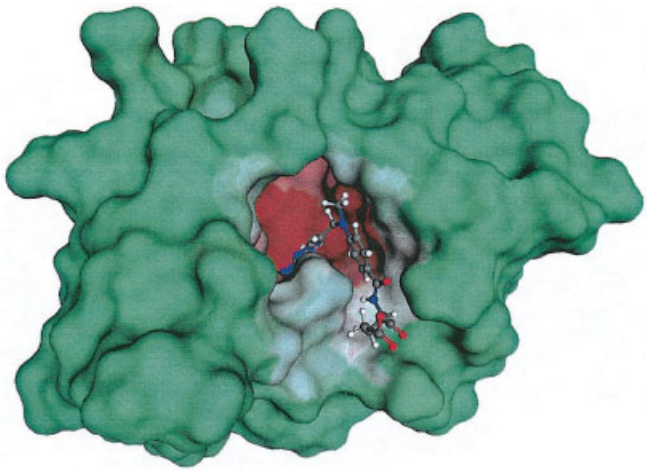

Figure 8. Binding site of the dihydrofolate reductase–methotrexate complex. The property describing the ability to form protein–ligand complexes is shown color-coded on the molecular surface of dihydrofolate reductase (red, high probability for complex formation; green, low probability). Methotrexate is represented as a ball-and-stick model.

The first two examples are protein–protein complexes. The first is the thymidilate kinase dimer[69] and the other the enzyme–inhibitor complex of β-trypsin with the bovine pancreatic trypsin inhibitor (BPTI).[70] The third and fourth examples are a protein–DNA complex, the p53 tumor suppressor complex,[71] and a protein–ligand complex, the enzyme–inhibitor complex of dihydrofolate reductase with methotrexate.[72] It can be seen that in the first example the binding site, shown in red, is identified correctly on the molecular surface, even if a small part (yellow circle) is characterized as nonbonding. All binding sites of the other examples are predicted perfectly. Besides the correct one, the NN identifies a small number of additional binding sites on the molecular surface. But, these are always smaller than the correct one and are mostly composed of only one domain. As mentioned above, it is unlikely that these isolated domains build stable complexes.

Conclusion

In this article, we presented an algorithm for the identification of possible binding sites of biomolecular complexes based only on the 3D structure of one complex partner, the protein. The algorithm uses the segmentation of the molecular surface into overlapping domains, described in the first article of this series.[1] The mean values of six different molecular properties are calculated for these domains, which are then used as input to a neural network. After the training with known complexes from the PDB,[3,4] the network was able to classify the surface domains of all proteins included in this database into four classes according to their ability to form complexes with other proteins, DNA/RNA molecules, and small ligands or cannot build complexes at all with an identification rate of 76%. If the second best prediction of the NN is taken into account, the identification rate increases to 90%. This information was mapped back onto the molecular surface of four representative example complexes. In all four examples, the correct binding site could be identified as the largest region of high ability to form a complex of the specific complex class. Besides the correct one, the NN identifies a small number of additional binding sites on the molecular surface. But, these are too small to build stable complexes. For this reason, only large binding sites found by the algorithm must be considered in the investigations if an unknown complex is treated with the method described here. On the other hand, if additional large binding sites are found this could give hints for the identification of alternative binding mechanisms, like the noncompetitive inhibition of enzymes, and the specific development of drugs, binding to these alternative sites.

At the moment, the identification of the binding site is done by a visual inspection of the ability to form a complex mapped on the molecular surface. The next step in the ongoing development of this method will be the automatic identification of large clusters of domains with a high probability for complex formation representing possible binding sites. These binding sites can then be used in a large variety of applications, e.g., a number of often used docking programs, like DOCK[73] and FlexX,[74] are based on the optimization of ligand orientation and conformation in the active site of the protein, the position of which must be known in advance. Therefore, they rely on already known complex structures of the same enzyme with different inhibitors or on other experimental or computational methods to determine the location of the active site. Our method could be used as a fast preprocessing step for these docking algorithms so that a manual identification of the active site is no longer necessary. Most steps required for the identification of the active site are needed for the docking algorithms anyway and could be done by the automatic procedure described in this work.

Further improvements are also possible for our method. It should be tested if other molecular properties than those used here can improve the identification rate of the NN. A large number of descriptors are proposed in the literature that have successfully been used in quantitative structure–activity relationships.[75–80] A number of these descriptors can be adapted to our method. With the usage of additional parameters, the optimization of the NN must be going hand in hand. This optimization can include the variation of the network architecture but also the use of different network types, like Kohonen networks.[59,81]

An improvement of the algorithm could also be possible with an optimized training set for the NN. A number of structures in the PDB include only a part of the molecules involved in the complex formation. The reason for this is that the whole biomolecular systems are often too complex to be crystallized. A large part of the database is also built out of free enzyme structures, where the enzyme is not bonded to a substrate or inhibitor. It is important that these structures are not included in the training set because the automatic procedure would classify the binding domains of these proteins as nonbonding and the NN would be trained to assign the complex class "nonbonding" to these and other domains with similar properties. It is also possible to divide the complexes of the training set into additional classes. On the one hand, the protein–protein complexes in the PDB are mainly oligomers and enzyme–inhibitor complexes. While most interfaces of oligomers are flat and lipophilic, the active sites of enzymes are designed to catalyze the reactions of substrates, which can be small. Therefore, the binding sites of these enzymes are more like the ones of proteins binding small molecules, which we included in the protein–ligand complex class in this work, and the inhibitors try to mimic the properties of the small molecules. A separation of these different types in two classes or the inclusion of the enzyme–inhibitor complexes in the protein–ligand class could be helpful. On the other hand, the protein–ligand class is diverse. It includes enzymes with small inhibitors but also complexes with carbohydrates and many other molecules important for the structural features and bioactivity of the complexes. By dividing the protein–ligand class according to the function of the ligand, properties specific for these particular functions could be identified and used in the further development of the neural network.

Acknowledgments

This work was supported by the Fonds der Chemischen Industrie, Frankfurt, the Deutsche Forschungsgesellschaft, Bonn, and the Land Hessen. T.E.E. gratefully acknowledges the Alexander von Humboldt Foundation for a Feodor Lynen research fellowship.

References

1. Exner, T. E.; Keil, M.; Brickmann, J. J Comput Chem 2002, 23, 1176–1187.
2. Exner, T. E.; Keil, M.; Brickmann, J. J Comput Chem 2002, 23, 1188–1197.
3. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E. Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112, 535.
4. Berman, H. M.; Westbrook, J.; Feng, Z.; Filliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
5. Elcock, A. H.; Sept, D.; McCammon, J. A. J Phys Chem B 2001, 105, 1504.
6. Chothia, C.; Janin, J. Nature 1975, 256 705.
7. Janin, J.; Miller, S.; Chothia, C. J Mol Biol 1988, 204, 155.
8. Janin, J.; Chothia, C. J Biol Chem 1990, 265, 16027.
9. Jones, S.; Thornton, J. M. Progr Biophys Mol Biol 1995, 63, 31.
10. Jones, S.; Thornton, J. M. Proc Natl Acad Sci USA 1996, 93, 13.
11. Lin, S. L.; Nussinov, R. J Mol Graph 1996, 14, 78.
12. Jones, S.; Thornton, J. M. J Mol Biol 1997, 272, 121.
13. Jones, S.; Thornton, J. M. J Mol Biol 1997, 272, 133.
14. Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Protein Sci 1997, 6, 53.
15. Bogan, A. A.; Thorn, K. S. J Mol Biol 1998, 280, 1.
16. Conte, L. L.; Chothia, C.; Janin, J. J Mol Biol 1999, 285, 2177.
17. Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben Tal, N. Proteins 2001, 43, 89.
18. McCoy, A. J.; Chandana, E.; Colman, P. M. J Mol Biol 1997, 268, 570.
19. Xu, D.; Tsai, C. J.; Nussinov, R. Protein Eng 1997, 10, 999.
20. Xu, D.; Lin, S. L.; Nussinov, R. J Mol Biol 1997, 265, 68.
21. Young, L.; Jernigan, R. L.; Covell, D. G. Protein Sci 1994, 3, 717.
22. Scarsi, M.; Majeux, N.; Caflisch, A. Proteins 1999, 37, 565.
23. Pacios, L. F. J Chem Info Comput Sci 2001, 41, 1427.
24. Jones, S.; van Heyningen, P.; Berman, H. M.; Thornton, J. M. J Mol Biol 1999, 287, 877.
25. Wang, R.; Liu, L.; Lai, L.; Ben Tal, N.; Tang, Y. J Mol Model 1998, 4, 379.
26. Brady, G. P.; Stouten, P. F. W. J Comput-Aided Mol Design 2000, 14, 383.
27. Danziger, D. J.; Dean, P. M. Proc Roy Soc Lond B Biol 1989, 236, 101.
28. Peters, K. P.; Fauck, J.; Frömmel, C. J Mol Biol 1996, 256, 201.
29. Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. J Mol Biol 2001, 307, 841.
30. Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein Sci 1996, 5, 2438.
31. Ruppert, J.; Welch, W.; Jain, A. N. Protein Sci 1997, 6, 524.
32. Levitt, D. G.; Banaszak, L. J. J Mol Graph 1992, 10, 229.
33. Delaney, J. S. J Mol Graph 1992, 10, 174.
34. Laskowski, R. A. J Mol Graph 1995, 13, 323.
35. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. J Comput Chem 1992, 13, 505.
36. Exner, T. E.; Keil, M.; Moeckel, G.; Brickmann, J. J Mol Model 1998, 4, 343.
37. Edelsbrunner, H.; Mücke, E. P. ACM Transact Graph 1994, 13, 43.
38. Goodford, P. J. J Med Chem 1985, 28, 849.
39. Kellogg, G. E.; Semus, S. F.; Abraham, D. J. J Comput-Aided Mol Design 1991, 5, 545.
40. Klebe, G. J Mol Biol 1994, 237, 212.
41. Laskowski, R. A.; Thornton, J. M.; Humblet, C.; Singh, J. J Mol Biol 1996, 259, 175.
42. Hendlich, M.; Rippmann, F.; Barnickel, G. J Mol Graph Model 1997, 1997, 15(6), 359.
43. MacCallum, R. M.; Martin, A. C.; Thornton, J. M. J Mol Biol 1996, 262, 732.
44. RCSB Protein Data Bank; http://www.rcsb.org/pdb/.
45. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
46. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379.
47. Richards, F. M. Annu Rev Biophys Biol 1977, 6, 151.
48. Connolly, M. L. Science 1983, 221, 709.
49. Connolly, M. L. J Appl Crystallogr 1983, 16, 548.
50. Brickmann, J.; Heiden, W.; Vollhardt, H.; Zachmann, C. D. New man–machine communication strategies in molecular modelling; In: Hunter, L.; Shriver, B. D., eds. Proceedings of the 28th Annual Hawaii International Conference on System Sciences, Biotechnology Computing, vol. V; IEEE Computer Society Press: Los Alamitos, CA, 1995; p 273–282.
51. Brickmann, J.; Goetze, T.; Heiden, W.; Moeckel, G.; Reiling, S.; Vollhardt, H.; Zachmann, C. D. Interactive visualization of molecular scenarios with MOLCAD/SYBYL. In: Bowie, J. E., ed. Data Visualization in Molecular Science: Tools for Insight and Innovation; Addison-Wesley: Reading, MA, 1995; p 83–97.
52. Brickmann, J.; Exner, T.; Keil, M.; Marhöfer, R.; Moeckel, G. Molecular models: Visualization. In: Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R., eds. The Encyclopedia of Computational Chemistry, vol. 3; John Wiley & Sons: Chichester, UK, 1998; p 1679–1693.
53. Keil, M.; Brickmann, J. In preparation.
54. Ghose, A. K.; Crippen, G. M. J Comput Chem 1986, 7, 565.
55. Ghose, A. K.; Pritchett, A.; Crippen, G. M. J Comput Chem 1988, 9, 80.
56. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. J Chem Info Comput Sci 1989, 29, 163.
57. Jäger, R.; Schmidt, F.; Schilling, B.; Brickmann, J. J Comput-Aided Mol Design 2000, 14, 631.
58. Keil, M. Modellierung und vorhersage von strukturen biomolekularer assoziate auf der basis von statistischen datenbankanalysen (dissertation); Darmstadt University of Technology: Darmstadt, Germany, 2002.
59. Zupan, J.; Gasteiger, J. Neuronal Networks for Chemists: An Introduction; VCH: Weinheim, Germany, 1993.
60. Schneider, G.; Wrede, P. Progr Biophys Mol Biol 1998, 70, 175.
61. Zupan, J. Neural networks in chemistry. In: Schleyer, P. v. R.; Allinger, N. C.; Clark, T.; Gasteiger, J.; Kollmann, P. A.; Schaefer, H. F. III; Schreiner, P. R., eds. The Encyclopedia of Computational Chemistry; John Wiley & Sons: Chichester, UK, 1998; p 1813–1827.
62. Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. J Comput-Aided Mol Design 1996, 10, 521.
63. Sadowski, J.; Kubinyi, H. J Med Chem 1998, 41, 3325.
64. Qian, N.; Sejnowski, T. J. J Mol Biol 1988, 202, 865.
65. Rost, B.; Sander, C. J Mol Biol 1993, 232, 584.
66. Reczko, M. SAR QSAR Environ Res 1993, 1, 153.
67. Zell, A.; Mache, N.; Sommer, T.; Korb, T. Design of the SNNS neural network simulator. In: Östereichische Artificial-Intelligence-Tagung; Springer-Verlag: Wien, Germany, 1991; p 93–102.
68. Rumelhart, D. E.; McClelland, J. L. Parallel Distributed Processing: Explorations in the Microstructure of Cognition; MIT Press: Cambridge, MA, 1986.

69. Lavie, A.; Vetter, I. R.; Konrad, M.; Goody, R. S.; Reinstein, J.; Schlichting, I. Nat Struct Biol 1997, 4, 601.
70. Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. Acta Crystallogr B 1983, 39, 480.
71. Cho, Y.; Gorina, S.; Jeffrey, P. D.; Pavletich, N. P. Science 1994, 265, 346.
72. Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. J Biol Chem 1982, 257, 13650.
73. Kuntz, I. D.; Blaney, J. M.; Oatlen, S. J.; Langridge, R.; Ferrin, T. E. J Mol Biol 1982, 161, 269.
74. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. J Mol Biol 1996, 261, 470.
75. Gupta, S. P. Progr Drug Res 2001, 56, 121.
76. Katritzky, A. R.; Petruhkin, R.; Tatham, D.; Basak, S.; Benfenati, E.; Karelson, M.; Maran, U. J Chem Info Comput Sci 2002, 41, 679.
77. Kubinyi, H. Drug Discovery Today 2002, 2, 457.
78. Martin, Y. C. Perspect Drug Discovery Design 1998, 12–14, 3.
79. Warne, M. A.; Nicholson, J. K. Progr Environ Sci 1999, 1, 327.
80. Winkler, D. A. Brief Bioinformat 2002, 3, 73.
81. Kohonen, T. Biol Cybernet 1982, 43, 59.