From: ISMB-95 Proceedings. Copyright © 1995, AAAI (www.aaai.org). All rights reserved.

Protein Docking Combining Symbolic Descriptions of Molecular Surfaces and Grid–Based Scoring Functions *

F. Ackermann and G. Herrmann and F. Kummert and S. Posch and G. Sagerer
Bielefeld University
P.O.Box 100131
33501 Bielefeld
Germany
{friedric | grit | franz | posch | sagerer}@techfak.uni–bielefeld.de

D. Schomburg
GBF (Gesellschaft für Biotechnologische Forschung)
Mascheroderweg 1
38124 Braunschweig
Germany
schomburg@venus.gbf–braunschweig.d400.de

Abstract

With the growing number of known 3D protein structures, computing systems, that can predict where two protein molecules interact with each other is becoming of increasing interest. A system is presented, integrating preprocessing like the computation of molecular surfaces, segmentation, and searching for complementarity in the general framework of a pattern analyzing semantic network (ERNEST). The score of coarse symbolic computations is used by the problem independent control strategy of ERNEST to guide a more detailed analysis considering steric clash and judgements based on grid–based surface representations. Successfull examples of the docking system are discussed that compare well with other approaches.

Introduction

Life and development of all organisms are mainly determined by molecular interactions, e.g. between DNA and proteins, proteins and proteins, proteins and carbohydrates, proteins and small molecules, or proteins with membranes. Among these, protein–protein interactions play an especially important role, like in interactions between antibodies and antigens, receptors and peptide– or protein–hormones, enzymes and substrates or inhibitors. With the growing number of known protein 3D structures predicting whether and where protein molecules interact with each other is becoming of increasing interest. In the cell protein molecules associate spontaneously without the need of external assistance. They form a complex if they exhibit a high affinity to each other, i.e. the free energy of the complex is lower than that of the two single solvated protein molecules. So, given the right methods it should be possible to predict the docking site and orientation of two protein molecules.

With the rapid advance of molecular modelling techniques a number of attempts have been made to answer the question where two biological macromolecules, known to interact, bind. Although progress has been made, a generally applicable method has still to be developed. The simulation of molecular docking involves two principal steps:

1. Prediction of possible docking sites: In the analysis of complexes with known 3D structure it was found that protein complexes reveal a striking degree of spatial and electrostatic complementarity between the interacting surfaces. Therefore this step mainly consists of the detection of geometrically and electrostatically complementary surface regions.
2. Evaluation of the identified sites with respect to the free energy of interaction.

In the current paper our approach for the first step is

*This work has been sponsored by the German Ministry for Research and Technology by grant no. 01 IB 307 C.

introduced.

System Overview

Solving the protein docking problem requires, at last in principle, consideration of the complete 6D space of relative rotations and translations. Whereas this was indeed the strategy in some approaches (e.g. (Katchalski-Katzir *et al.* 1992)), using current technology and algorithms this space is too large to be sampled exhaustively. Therefore restrictions were formulated and applied, like interactively specifying a binding site in (Wodak & Janin 1978). Alternatively, in (Connolly 1986; Wang 1991; Norel *et al.* 1994) "critical points" were computed. Among these only small subsets were compared in search for the correct docking position.

Our approach is in the spirit of the second method: The semantic network ERNEST (Kummert *et al.* 1993; Niemann *et al.* 1990a) serves to represent and use symbolic and numerical knowledge about the protein docking problem. This symbolic description comprises a hierarchical model of protein–protein complexes, the involved protein surfaces and their chemical features. Starting by generating the molecular surface as a set of surface points, the next step is its segmentation into significant convex and deep concave regions. ERNEST searches for possible docking positions, trying to match the previously segmented regions, comparing their shapes and volumes. The score of this coarse symbolic comparison is used by the problem independent control strategy of ERNEST to guide a more detailed analysis considering steric clash and judgements based on the cross correlation of grid–based surface representations. Thereby, not only the geometrical fit in a finer resolution, but also the complementarity of chemical attributes is evaluated. This gives the final scoring of the docking positions.

A Semantic Network for Protein Docking

In this section we first describe the semantic network ERNEST, and subsequently sketch the declarative and procedural knowledge as modelled for the protein docking problem.

The semantic network system ERNEST

ERNEST is a semantic network system facilitating knowledge representation and utilization. As a rough simplification a semantic network is a graph containing nodes and directed links. Sometimes, such networks are viewed as a graphical representation of first order predicate calculus. But since the development of similar approaches like KL-ONE ((Brachman & Schmolze 1985)) or PSN ((Mylopoulos, Shibahara, & Tsotsos 1983)) the stronger expressive power of these networks is out of discussion. Especially, PSN showed the way to integrate a procedural semantics additionally to the declarative structures. In ERNEST, the problem independent procedural semantics of PSN is generalized to problem independent inference rules which are comparable to the resolution rule in predicate logic. Furthermore, it was possible to define a problem independent control alorithm which substitutes backtracking as known for PROLOG. Therefore ERNEST defines a complete knowledge representation system.

To achieve a declarative and procedural semantics only three different types of nodes and three different types of links suffice. The first type of nodes, *concepts*, represents classes of objects, events, or abstract conceptions with some common properties. In the context of pattern understanding the goal is the interpretation of the sensor signal in terms of these concepts modeled in the knowledge base. The second node type, called *instance*, represents these extentions of a concept. An instance is a copy of the related concept except that the general description is substituted by concrete values calculated from the signal data. In an intermediate state of processing it may occur, that instances to some concepts cannot be computed because certain prerequisites are missing. Nevertheless, the available information can be used to constrain an uninstantiated concept. This is done with the node type *modified concept* representing modifications of a concept due to intermediate results of the analysis.

The three link types *part*, *specialization*, and *concrete* are also complex data structures characterizing the properties of the link. Via *parts* a concept is decomposed into its natural components. The link type *specialization* connects a concept with a more general concept. Closely related to that link is an inheritance mechanism by which a special concept inherits all properties ot its general ones. For a clear distinction of knowledge of different levels of abstraction the link type *concrete* is introduced. In the definition of a concept, there may be parts or concretes which are *obligatory* and others which are *optional*. A set of parts and concretes, either obligatory or optional, is called a *modality set* describing a particular quality of the conception. In addition to its links, a concept is described by attributes representing its features and restrictions for these values according to the modeled knowledge. Furthermore, relations defining constraints for the attributes and parts can be specified and must be satisfied for valid instances.

The main activity during an analysis process is the computation of instances given certain sensor data. This aspect, the utilization of the represented knowl-

edge, is defined by six rules and is the basis for a problem independent control algorithm. For example, the creation of instances is based on the fact that recognition of a complex object requires all its parts as a prerequiste. Since the results of an initial segmentation are not perfect, the definition of a concept is completed by a judgement function estimating the degree of correspondence of a signal area to the term defined by the related concept. On the basis of these estimates and the inference rules, an A*–like control algorithm is applied see (Nilsson 1971). For a detailed description of ERNEST see (Niemann *et al.* 1990a; Kummert *et al.* 1993). This semantic network system has been applied to various signal interpretation tasks, including speech understanding, interpretation of industrial scenes, and diagnostic interpretation of image sequences of the heart (see (Kummert *et al.* 1993; Niemann *et al.* 1990b)).

An application of protein docking in ERNEST

The design of a ERNEST network for a given application has some resemblance with the approach of object oriented programming. First, the conceptions to be represented as concepts and the links between them are specified. Next, their attributes and relations between attributes are decided upon and finally the procedures to compute attributes, relations and judgements have to be realized.

Therefore, the first step is to symbolically model the necessary declarative knowledge for the docking problem (see figure 1). In the network, proteins are described by surface regions relevant for docking with appropriate attributes for their geometrical and chemical features. Considering geometric features like shape and the mean hydrophobicity of a region, the following concepts are defined: "concave edge" (CONC_EDG), "convex edge" (CONV_EDG), "flat region" (FLAT_REG), "hydrophobic pocket" (H_PHOB_PO), "hydrohpil pocket" (H_PHIL_PO), "hydrophobic arm" (H_PHOB_AR), and "hydrophil arm" (H_PHIL_AR). The concepts CONC_EDG, CONV_EDG, and FLAT_REG are large regions and are decomposed into "hydrophobic groove"(H_PHOB_GR), "Hydrophil groove" (H_PHIL_GR), "hydrophobic small concave edge" (H_PHOB_SCCE), "hydrophil small concave edge" (H_PHIL_SCCE), "hydrophobic small convex edge" (H_PHOB_SCVE), "hydrophil small convex edge" (H_PHIL_SCVE) , "hydrophobic roof" (H_PHOB_RO), and "hydrophil roof" (H_PHIL_RO). With concrete links all these concepts are related to appropriate concepts in the abstraction level of segmentation. At this level the results of our segmentation algorithm are incorporated into the semantic network.

The concept DOCKING_SITE defines a match of two possible regions with complementary shape and hydrophobicity, thus describing potential docking positions taking only local information into account. To this end, various modality sets are defined for this concept. Each has two obligatory parts describing the two regions which are required to define a possible coarse docking constallation (e.g. CONC_EDG of protein 1 and CONV_EDG of protein 2). For a better fit we consider also small regions as parts of the large ones. Hence each modality set of DOCKING_SITE has an optional part LOC_DOCK, which also is modelled with several modality sets. Again, these modalities have two obligatory parts of two complementary small regions. For example, the "concave edge" of protein 1 can optionally encloses a "hydrophil small convex edge", and the "convex edge" on protein 2 can optionally encloses a "hydrophil small concave edge". Using this information, the potential coarse docking position can be determined more precisely considering local values of hydrophobicity and shape. The actually used segmentation technique allows not an evaluation of complementarity for cases of saddle interactions and ridge and valley interactions. Enhanced segmentation algorithms of protein surfaces will be implemented.

To instantiate the concept COMPLEX, instances for the obilgatory parts PROTEIN_1, PROTEIN_2, and DOCKING_SITE must exist. Instances for PROTEIN_1 and PROTEIN_2 are found instantiating their optional parts, the segmented surface regions. Furthermore, instances of DOCKING_SITE for all possible combinations of complementary regions are generated, described by the defined modality sets. During analysis these instances are judged by shape complementarity, hydrophobicity complementarity and size of the docking area. Due to the A^*–driven control, the best judged instances of DOCKING_SITE are selected for further processing. During instantiation of the concept COMPLEX, the steric clash is assessed by a related judgement function. For the current system, a grid with 20^3 points is used and the accepted overlap of the two proteins is confined to a specified interval enforcing a minimal overlap and preventing a deep molecular penetration. Thus, also global criteria of the proposed docking position are considered. With a grid-based technique a final scoring is computed to rate geometrical and chemical complementarity on a finer scale and to fine tune the rotation and translation of the coarse docking position.

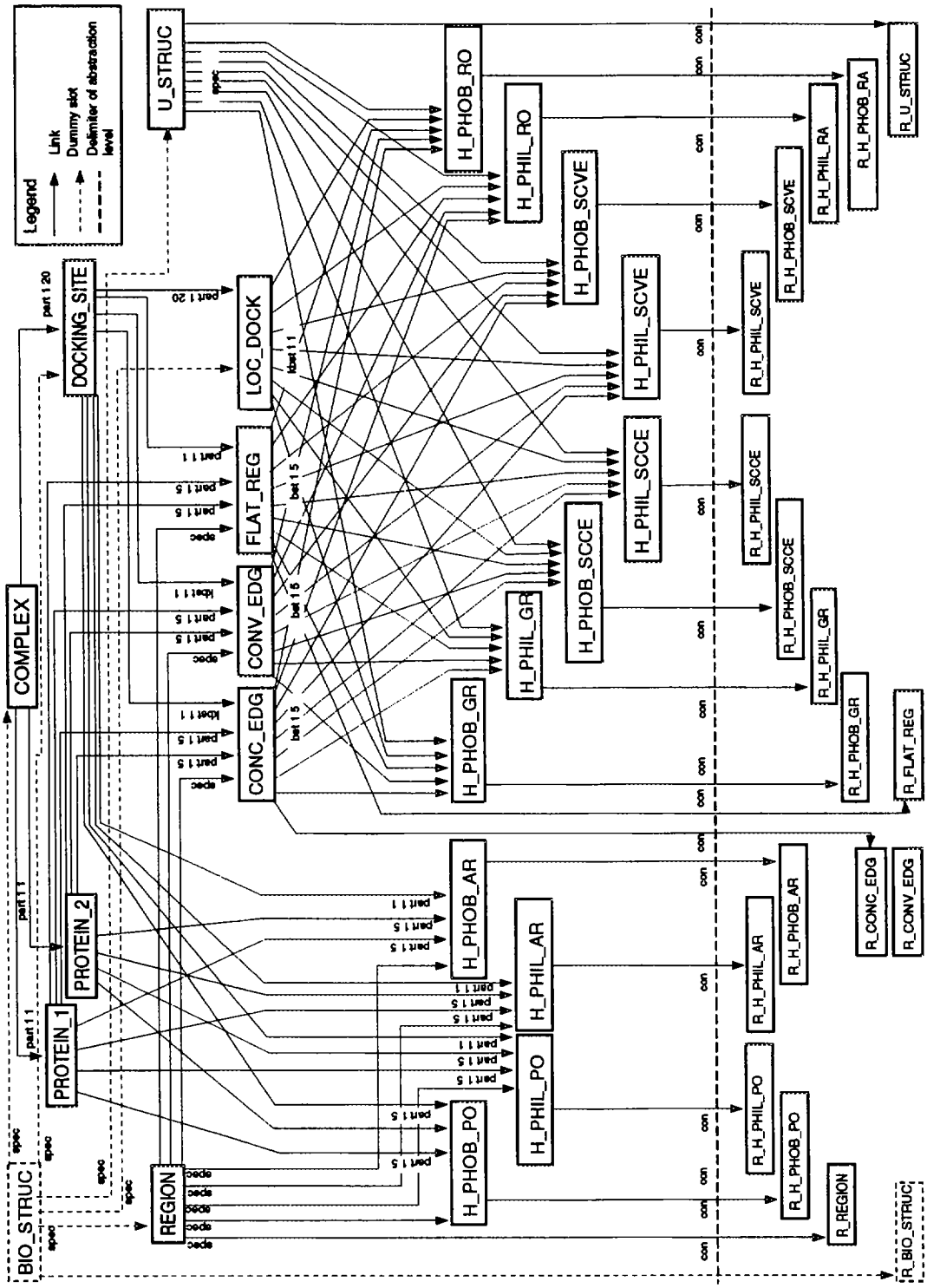

Figure 1: Model of the semantic network application for protein docking

Grid–based scoring of geometrical and chemical complementarity

As discussed in the previous section, the analysis of two given molecular surfaces results in a set of instances of the concept complex. Each instance proposes a discrete docking position (DDP), defined by the translational vector and the rotation about its center of mass of the second subunit relative to the first one.

In this section we describe the grid–based scoring of a DDP modifying and enhancing the method of (Katchalski-Katzir *et al.* 1992). We proceed as follows: For each DDP first a bounding box is calculated, called *clipping box* CB. It contains the volume, where the two proteins touch each other, given the proposed transformation. In this clipping region a discrete sampling of geometrical and chemical protein features is performed. Subsequently the corresponding discrete samplings are crosscorrelated, resulting in two components of a two dimensional scoring vector. In order to achieve a fine tuning of the DDP, this correlation is repeated for a small set of up to nine nonredundant rotations lying in a narrow cone around the DDP.

The main advantages of this techique are twofold: On one hand it realizes a hierarchical approach, computing the cross correlation only for a region of interest, namely the clipping box. Thus, the grids we use have only 16^3 up to 32^3 grid points, reducing computational complexity by a factor $O(10) - O(100)$ compared to (Katchalski-Katzir *et al.* 1992). On the other hand we avoid a complete sampling of the rotational parameter space, which usually is the major draw back of grid based evaluations of protein docking positions (Wang 1991; Katchalski-Katzir *et al.* 1992).

Sampling of molecular surface attributes in 3D clipping boxes

Given a DDP proposed by the docking net, we translate and rotate the second molecule accordingly. The resulting situation is sketched in figure 2: The two proteins are located near each other, usually penetrating to some extend due to inaccuracy of the proposed DDP. (The conditions imposed by the steric clash test for DDPs exclude the case of nontouching proteins, see last section.) Now the minimal and maximal coordinates $(x_{h,i}^{min}, x_{h,j}^{min}, x_{h,k}^{min})$ and $(x_{h,i}^{max}, x_{h,j}^{max}, x_{h,k}^{max})$ for both subunits $h = 1$ and $h = 2$ are calculated and sorted for each component. The two extremes are discarded and the two remaining ones give the lower and the upper bound of the clipping region for each component. If the volume VCB of the clipping region exceeds a threshold $VCB_c = 10,000\mathring{A}^3$ a grid with 32^3 grid points is choosen for the sampling, 16^3 points otherwise.

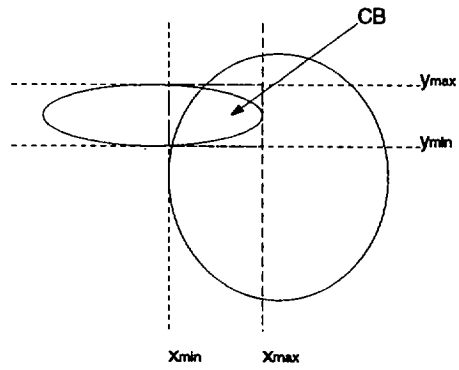

Figure 2: 2D sketch of the calculation of the clipping box CB: Minimal and maximal coordinates of the proteins in each direction are computed. The four values of each direction are sorted and the largest and the smallest one are discarded. The two in the middle give the minimal and maximal coordinates of CB in this direction.

As a next step each subunit is sampled within the appropriate grid, neglecting all atoms outside CB. Each grid point $\underline{x}_{ijk}$ takes two values:

(1) The geometry is reflected by a *surface function* S_P representing the solvent–accessible surface of protein P according to (Richards 1977):

$$S(\underline{x}_{ijk}) = \begin{cases} 0, & \Longleftrightarrow \quad \forall\, \kappa \in P \mid d(\kappa, \underline{x}_{ijk}) > D(\kappa) \\ 1, & \Longleftrightarrow \quad \exists \kappa \in P \mid d(\kappa, \underline{x}_{ijk}) \leq D(\kappa) \wedge \\ & \quad S(\underline{x}_{i'j'k'}) = 0 \text{ for at least} \\ & \quad e_c \text{ neighbours} \\ \rho_P, & \Longleftrightarrow \quad \exists \kappa \in P \mid d(\kappa, \underline{x}_{ijk}) \leq D(\kappa) \wedge \\ & \quad S(\underline{x}_{i'j'k'}) \neq 0 \text{ for at least} \\ & \quad 27 - e_c \text{ neighbours} \end{cases}$$

(Eq. 1) Hereby, grid points lying on the surface are labeled with 1, points outside the protein are set to zero and interior points get the value ρ_P. $D(\kappa)$ is the sum of the van der Waals radius of atom κ (taken from (Weiner *et al.* 1984)) and the radius of a water sized probe sphere (1.4Å). The threshold e_c affects the thickness of the surface layer, we always use $e_c = 2$. ρ_P is normally a small positive value in the interior of one protein and a negative value in the interior of the other, thus generating a penalty term for intramolecular penetration correlating S_1 and S_2.

(2) Driving force for the formation of complexes is the gain of free energy. Several interatomic forces and entropic factors related to water and the proteins themselves contribute to it (Banaszak, Birktoft, & Barry

1981). Geometrical complementarity is not only a precondition for some, but rather results directly in the exclusion of water from the binding site and therefore in an increase of entropy. This gain in entropy will be especially large, if hydrophobic residues are involved. Therefore each grid point at the surface is labeled with a second value, the *hydrophobicity function* H, describing the hydrophobic nature of the residue constituting it:

$$H(\underline{x}_{ijk}) = \begin{cases} 0, & \Longleftrightarrow \ \forall \kappa \in P \mid d(\kappa, \underline{x}_{ijk}) > D(\kappa) \\ max(h(\kappa), \bar{h}), & \Longleftrightarrow \ \exists \kappa \in P \mid d(\kappa, \underline{x}_{ijk}) \leq D(\kappa) \wedge \\ & \quad S(\underline{x}_{i'j'k'}) = 0 \text{ for at least} \\ & \quad e_c \text{ neighbours} \\ 0, & \Longleftrightarrow \ \exists \kappa \in P \mid d(\kappa, \underline{x}_{ijk}) \leq D(\kappa) \wedge \\ & \quad S(\underline{x}_{i'j'k'}) \neq 0 \text{ for at least} \\ & \quad 27 - e_c \text{ neighbours,} \end{cases}$$

(Eq. 2) where $h(\kappa)$ is identical with the hydrophobicity of the residue to which atom κ belongs to and $\bar{h} = -0.49$. The hydrophobicity of residues is taken from the consensus scale (Kyte & Doolittle 1982). Only the hydrophobic moments exceeding the mean $\bar{h}$ are retained, since a contribution to the scoring vector is only intended if both residues in contact are hydrophobic. If only one of them is hydrophobic, further consideration e.g. inspection of the signs of polarities is necessary.

Final scoring and sorting of results

Once the two molecules are sampled, the cross correlation

$$c_S(I,J,K) = \sum_{i,j,k} S_1(\underline{x}_{i,j,k}) \cdot S_2(\underline{x}_{i+I,j+J,k+K})$$

$$c_H(I,J,K) = \sum_{i,j,k} H_1(\underline{x}_{i,j,k}) \cdot H_2(\underline{x}_{i+I,j+J,k+K}) \quad (3)$$

on the grid is calculated in each of the two components making use of the convolution theorem to reduce the complexity from $O((N^3)^2)$ to $O((N^3)\mathrm{ld}N^3)$ (Press *et al.* 1992). As outlined in the last section, this is iterated for each rotation of the local sampling of the rotational parameter space.

The cross correlation of the first component becomes large iff the surfaces exactly fit each other and no penalty for penetration is produced by the *DDP*. The cross correlation of the second component in addition weights the hydrophobicity of the matched surface parts.

Both contributions have to be considered for the final scoring. Instead of deriving an absolute score, almost always leading to the problem of relative weightings, an ordering of the proposed docking positions is computed according to the following order relation (Eq. 4):

$$DDP_1 \left\{ \begin{array}{c} > \\ < \\ = \end{array} \right\} DDP_2 \Longleftrightarrow$$

$$\begin{cases} c_{S1} - c_{S2} > V_S & \vee & |c_{S1} - c_{S2}| \leq V_S & \wedge \\ & & c_{H1} - c_{H2} > V_H & \\ c_{S1} - c_{S2} < -V_S & \vee & |c_{S1} - c_{S2}| \leq V_S & \wedge \\ & & c_{H1} - c_{H2} < -V_H & \\ |c_{S1} - c_{S2}| \leq V_S & \wedge & |c_{H1} - c_{H2}| \leq V_H. & \end{cases}$$

Thus, *DDP*s are lexicographically ordered, using intervalls to test for equality. The size of those intervalls is determined by the thresholds V_S and V_H. (For example $V_S = \infty$ results in completely neglecting the first component for the sorting, because all values are considered to be equal.)

Results

Complete tests were carried out for the three protein–protein complexes Uteroglobin (PDB–Id 2utg), trypsinogen complex with porcine pancreatic secretory trypsin inhibitor (PDB–Id 1tgs), and horse–liver alcohol dehydrogenase complex with NAD and DMSO (PDB–Id 6adh). For the last one, we aim at docking the two monomers against each other. The docked subunits have between 56 and 374 residues, thus covering a wide range of different structures.

As described, first the molecular surfaces are computed. For our experiments we use a grid based sampling of the proteins with isotropic lattice constant 1.0 Å. This results in sets of 2939 (trypsin inhibitor) up to 11860 surface points (first subunit of alcohol dehydrogenase).

The segmentation of molecular surfaces into their large regions is an enhancement of the method of (Lee & Rose 1985): First the convex hull of the point set is computed, using the program QuickHull (Barber, Dobkin, & Huhdanpaa 1993). Now, the intrinsic features of the convex hull are exploited to detect convexities (R_CONV_EDG), and subsequently to identify deep concave regions (R_CONC_EDG). The main ideas are:

(1) A R_CONC_EDG is an area that protrudes from the surrounding surface. Its tip lies directly on the surface of the convex hull (think of an irregularly formed object lying on a desk: Its tips are in direct contact with the table–top).

(2) A R_CONV_EDG is covered by large facets of the surface of the convex hull.

The results of the segmentation for the three complexes are summarized in table 1.

PDB–Id	# CONC	# CONV	Comparison of curvature	Comparison of size	Steric clash
2utg_1	9	2	45	6	2
2utg_2	6	3			
1tgs_1	10	5	60	3	2
1tgs_2	6	3			
6adh_1	13	5	163	54	38
6adh_2	11	6			

Table 1: Numbers of regions (large convexities and concavities) for the 1st and the 2nd subunit, respectively. The next two columns show the number of possible combinations, that fullfill the criteria applied to compare curvature and of size, respectively. The following column shows the number of pairs of regions that pass the steric clash test.

Smaller regions are identified by a statistical approach: For each surface point a vector of geometrical and chemical features is computed (solid angle, hydrophobicity according to different empirical scales). Based on these feature vectors, the surface points are classified with a vector quantizer.

Whereas both segmentation techniques are implemented and work succesfully on numerous examples, we use only the first one in the current implementation of the docking network.

Once ERNEST has instantiated the surface regions, the next step is to instantiate DOCKING_SITES matching these regions. These instances are judged solely considering a rough comparison of shape and size of the involved regions. Table 1 shows the decrease of possible docking positions considered for further processing due to this judgement. During the check for steric clash, the first time a true geometrical transformation of the second protein relative to the first has to be computed. The steric clash test eliminates 77% up to 97% of the initial positions. The remaining ones are considered as *DDP*s. It is worthwile to note, that in all cases the remaining solutions contain at least one that matches two regions of the correct docking site[1]. In other words, in all cases at least one *DDP* passing the steric clash test represents a translation and rotation in the neighborhood of the correct one. However, generally it will differ from the correct one by a certain amount due to the inacurracy of the preceeding rough calculations. Table 2 shows in its first column the minimal difference RMS between the solutions proposed for further processing and the correct one.

Beside this, in some cases the solution with minimal distance to the correct one is not the best–judged (of course, we make *no use* of knowledge of the correct solution during computation!). This is the case for 6adh in our examples. Therefore grid–based scoring of docking positions is necessary.

Column three and four of table 2 show the improvements, that were achieved by the grid based scoring, if the hydrophobicity is taken into account. This is an interesting result: In the proposed environment it is not sufficient to look at geometric complementarity alone, rather scoring using chemical *and* geometrical features is appropriate.

The obtained results may first be compared with the original approach from (Katchalski-Katzir *et al.* 1992). Whereas no RMS is given in that paper, we conclude from the stepsize of the angular sampling (20°) and the published lattice constant, that the accuracy of the results should be comparable to ours. To our opinion, there are two main improvements: First, our method does not require the known solution lying in the set of possible rotations, generated by angular sampling. Second a pronounced speed up was achieved ((Katchalski-Katzir *et al.* 1992) give 7.5hr on a Convex C–220 as CPU–time), mainly caused by the reduced grid sizes and only partial sampling of the rotational parameter space.

In (Norel *et al.* 1994) detailed data about results obtained with methods related to ours can be found. The discussed examples share one common complex (1tgs). The best RMS–deviaton in (Norel *et al.* 1994) is slightly better than ours, but it is one solution out of 144438 (with rank 552), whereas we give the final RMS–deviation of the best judged solution. There is a clear trade–off between the accuracy and the number of final solutions, that must be keeped in mind, if docking–results are compared.

In figure 3 the correct and best judged docking position of the trypsinogen with pancreatic secretory trypsin inhibitor are displayed graphically.

[1]There may be several such solutions since the area of close contact may consist of more than one segmented region.

PDB-Id	RMS_{min} after steric clash	rank of sol. with RMS_{min}	final RMS best-judged sol.		CPU-time [sec]
			$V_S = 0$	$V_S = 400.0$	
2utg	4.64	1	19.35	5.49	294.5
1tgs	2.50	1	2.87	1.78	71.1
6adh	7.76	6	7.1	6.97	1050.2

Table 2: The first column shows the accuracy of results that can be achieved by symbolic judgements alone. However, the correct solution is generally not ranked first after this step (col. 2). The third and fourth column show the accuracy of the best judged solution after checking for geometrical (col. 3) and geometrical *and* chemical complementarity (col. 4), respectively. Cf. eq. (4) for the meaning of V_S.
Total CPU-times on a DEC 3000 AXP 300 are given.

PDB-Id	Δx_1^{min}	Δx_2^{min}	Δx_3^{min}	Δx_1^{max}	Δx_2^{max}	Δx_3^{max}
2utg	0.89	0.47	0.84	1.32	1.27	1.21
1tgs	0.99	1.40	0.49	1.50	1.54	0.98
6adh	0.58	0.53	0.62	1.38	1.21	1.66

Table 3: The lattice constants in the grids are determined dynamically depending on the volume of CB. The table shows examples of minimal and maximal lattice constants in Å in either direction.

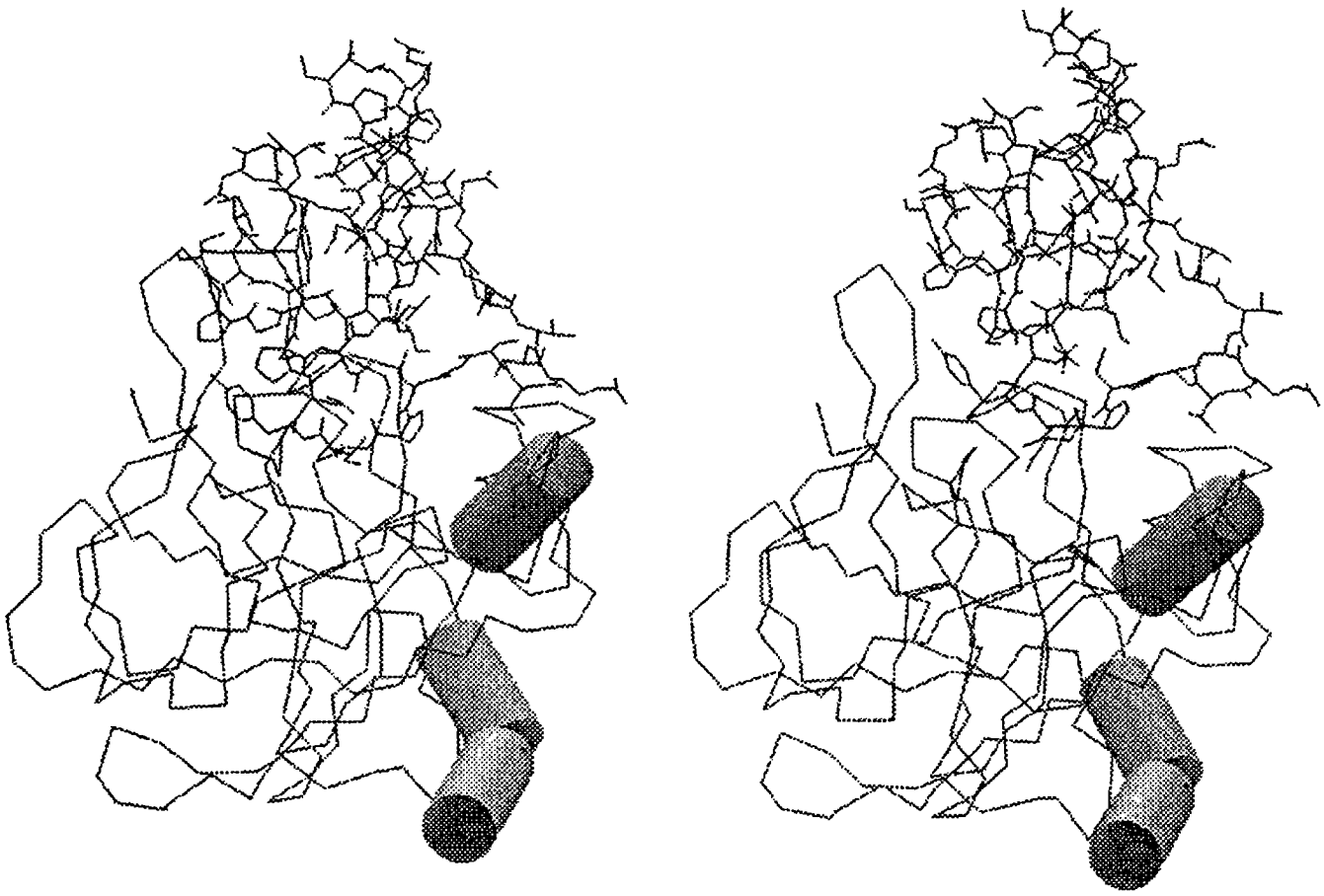

Figure 3: On the left the cocrystallized complex trypsinogen with pancreatic secretory trypsin inhibitor, PDB-ID 1tgs, is shown. On the right, the best judged solution as calculated by the docking system.

Conclusion

As outlined in the introduction, the need for computing systems, that can predict whether and where two given proteins with known 3D structure will dock, is of increasing importance. We introduced a system computing potential docking positions, integrating preprocessing like the computation of molecular surfaces, segmentation, and searching for complementarity in the general framework of a pattern analyzing system. To this end, a symbolic model of the docking problem was developed integrating geometrical and chemical features and concepts. As could be shown, this early integration is a striking point, if false local optima of judgement functions should be avoided. One may see a chance for learning from each other between chemistry and computer science in this matter of fact.

The further developments will mainly consist in a larger number of test examples and enhancements to the knowledge base including judgment of intermediate results and the overall control of the search for docking positions. Additionally, we believe that our approach is suited to handle also flexible molecules due to the local characterization of molecules and local computations performed by the system. Beside this, a database for the purposes of protein–protein docking should be integrated to facilitate access to all data, that can easily be computed offline.

Acknowlegement

We would like to thank R. Meier and Ch. Schillo for implementation of segmentation programs and the referees for their helpful comments.

References

Banaszak, L.; Birktoft, J.; and Barry, C. 1981. Protein–protein interactions and protein structures. In Frieden, C., and Nichol, L., eds., *Protein–Protein Interactions*. New York etc.: Wiley. 31–128.

Barber, C. B.; Dobkin, D. P.; and Huhdanpaa, H. 1993. The quickhull algorithm for convex hull. TR GCG53, Geometry Center, Minneapolis.

Brachman, R. J., and Schmolze, J. G. 1985. An overview of the kl-one knowledge representation language. *Cognitive Science* 9:171–216.

Connolly, M. L. 1986. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers* 25:1229–1247.

Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; and Vakser, I. A. 1992. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* 89:2195–2199.

Kummert, F.; Niemann, H.; Prechtel, R.; and Sagerer, G. 1993. Control and explanation in a signal understanding environment. *Signal Processing* 32:111–145.

Kyte, J., and Doolittle, R. 1982. A simple method for displaying the hydropathic character of a protein. *Journ. Mol. Biol.* 157:105–132.

Lee, R. H., and Rose, G. D. 1985. Molecular recognition. I. Automatic identification of topographic surface features. *Biopolymers* 24:1613–1627.

Mylopoulos, J.; Shibahara, T.; and Tsotsos, J. K. 1983. Building knowledge based systems: The psn experience. In McCalla, G., and Cercone, N., eds., *Knowledge Representation*. IEEE Computer Magazine. 83–89.

Niemann, H.; Sagerer, G.; Schroeder, S.; and Kummert, F. 1990a. ERNEST: A semantic network system for pattern understanding. *IEEE Trans. PAMI* 12:9.

Niemann, H.; Brünig, H.; Salzbrunn, R.; and Schröder, S. 1990b. A knowledge-based vision system for industrial applications. *Machine Vision and Applications* 3:201–229.

Nilsson, N. J., ed. 1971. *Problem Solving Methods in Artificial Intelligence*. New York: McGraw-Hill.

Norel, R.; Lin, S.; Wolfson, H.; and Nussinov, R. 1994. Shape complementarity at protein–protein interfaces. *Biopolymers* 34:933–940.

Press, W.; Teukolsky, S.; Vetterling, W.; and Flannery, B. 1992. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge: Cambridge University Press, 2 edition.

Richards, F. M. 1977. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.* 6:151–176.

Wang, H. 1991. Grid–search molecular accessible surface algorithm for solving the protein docking problem. *Journ. Comp. Chemistry* 12:746–750.

Weiner, S.; Kollmann, P.; Case, D.; Chandra Singh, U.; Ghio, C.; Alagona, G.; Profeta, S.; and Weiner, P. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106:765–784.

Wodak, S. J., and Janin, J. 1978. Computer analysis of protein–protein interaction. *Journ. Mol. Biol.* 124:323–342.