New Contact Measures for the Protein Docking Problem

Hans-Peter Lenhof

Max-Planck-Institut für Informatik

66123 Saarbrücken, Germany

email: len@mpi-sb.mpg.de

Abstract

We have developed and implemented a parallel distributed algorithm for the rigid-body protein docking problem. The algorithm is based on a new fitness function for evaluating the surface matching of a given conformation. The fitness function is defined as the weighted sum of two contact measures, the *geometric contact measure* and the *chemical contact measure*. The geometric contact measure measures the "size" of the contact area of two molecules. It is a potential function that counts the "van der Waals contacts" between the atoms of the two molecules (the algorithm does not compute the Lennard-Jones potential). The chemical contact measure is also based on the "van der Waals contacts" principle: We consider all atom pairs that have a "van der Waals" contact, but instead of adding a constant for each pair (a, b) we add a "chemical weight" that depends on the atom pair (a, b). We tested our docking algorithm with a test set that contains the test examples of Norel et al. [NLWN94] and Fischer et al. [FLWN95] and compared the results of our docking algorithm with the results of Norel et al. [NLWN94, NLWN95], with the results of Fischer et al. [FLWN95] and with the results of Meyer et al. [MWS96]. In 32 of 35 test examples the best conformation with respect to the fitness function was an approximation of the real conformation.

1 Introduction

Docking reactions play an important role in a large number of biochemical processes. Although the mechanisms of docking reactions are not well known, two complementarity principles seem to be important for the recognition and binding of docking partners. The first principle is the *shape complementarity principle*: the shapes of the molecules that build a docking complex are (locally geometrically) complementary, that is, there is a good fit between the surfaces of the docking partners. The second complementarity principle is the *chemistry principle*. It states that there is a strong chemical "complementarity" (with respect to hydrogen bonds, electrostatic interactions, hydrophobicity and so on) between the sites of docking partners.

To appear in the First Annual International Conference on Computational Molecular Biology (RECOMB 97), Santa Fe, New Mexico, January 20-23, 1997

Most of the known approaches to protein docking make use of these two principles and formulate the problem as a 3D matching problem: Given two proteins A and B, compute all rigid motions of B, with A fixed, such that the resulting conformations match large "chemically complementary" parts of the surfaces of A and B with minimal penetration of B into the interior of A.

In the above *rigid body docking* problem, the strong assumption is made that the two proteins are rigid. Of course, proteins are not rigid. They have certain dynamics that may have a strong influence on their chemical reactivity. For example "hinge-bendings" (movement of relatively rigid substructures of molecules by rotations about common hinges) have been observed in some molecules (see [BH84, FM90, DNS+92, GC91, SVC+92]) and in some ligand-receptor bindings (see [WWF+91, RSW92, MSS+89]). But the rigid body docking is still an important problem as Norel et al. [NLWN95] state: ".... a rigid body technique which achieves an efficient and accurate matching of the surfaces ..... may constitute a first step in addressing the more general (flexible) problem". A typical approach to tackle the conformational flexibility problem is to decompose the proteins into rigid parts. The rigid parts are matched individually using algorithms for the rigid body docking. Then, the "local" solutions are tested for "global" consistency and fitness. See for example [DSD+86, LK92, SNW95].

Connolly [Con3] proposed the first geometry-based approach for the protein docking problem. In his approach the "Connolly" surfaces [Con1, Con2] of the molecules were represented by a discrete point set. Connolly's algorithm works as follows: for certain quadruples of "critical points" on the surface of A all "similar" quadruples of critical points on the surface of B are determined. For each pair of similar quadruples a transformation is computed that maps one quadruple onto the other. The transformation is applied on molecule B and the resulting conformation is judged with respect to Connolly's scoring function.

Wang [Wan91] changed the definition of "critical points" slightly. In Wang's algorithm only two-point sets are matched. The matching of a pair of two-point sets does not determine a rigid transformation, i.e., there is still a rotational degree of freedom. Hence for each matching pair of two-point sets a search has to be carried out for suitable rotations of B around the axes given by the two points on the surface of A.

Norel et al. [NLWN94, NLWN95] use pairs of critical points and the corresponding surface normals to compute surface point matches. The fast algorithm of Norel et al.

was able to match successfully all but one docking example out of a selected PDB test set. Other successful approaches that rely heavily on the surface normals of critical points have been published by Lin et al. [LNF+94] and Fischer et al. [FNN+93, FLWN95]. In these papers an elegant computer-vision based technique called Geometric-Hashing [LW88] is used to solve the matching problem. Fischer et al. [FLWN95] developed and implemented a suite of docking tools which contains a geometric docking algorithm and an energy evaluation routine for judging intermolecular van der Waals and electrostatic interactions.

Katchalski-Katzir et al. [KSE+92] proposed correlation techniques for solving the surface matching problem. Meyer et al. [MWS96] are developing a new correlation based algorithm which yields excellent docking results (see Section (4)). Their results have not yet been published.

Ackermann et al. [AHP+95] decompose the surfaces of the proteins into segments. Their algorithm searches for surface segments that have a correlation of geometry and hydrophobicity using a knowledge based semantic network (ERNEST). For other (geometric) protein docking techniques see [KCF84, KBO+82, HCT94, EKS+95, CDJ91, JK91, SK91, BM92, WS92].

In this paper we present a parallel distributed algorithm for the rigid protein docking problem that is based on a new fitness function $Fit(conf)$ for scoring the surface matching of a given conformation $conf$. The fitness function

$$\begin{aligned} Fit(conf) &= w_{geo} * GeoFit(conf) + \\ &\quad w_{chem} * ChemFit(conf) \end{aligned}$$

is the weighted sum of two new contact measures $GeoFit(conf)$ and $ChemFit(conf)$ for measuring the geometric and chemical "complementarity" of the surfaces. The geometric contact measure $GeoFit$ measures the "size" of the contact area of two molecules. The measure is a potential function that counts the "van der Waals contacts" between the atoms of the two molecules (the algorithm does not compute the Lennard-Jones potential). The chemical contact measure $ChemFit(conf)$ is also based on the "van der Waals contacts" principle: we consider all atom pairs that have a "van der Waals" contact, but instead of adding a constant for each pair (a, b) we add a "chemical weight" $ChemWeight(a, b)$ that depends on the atom pair (a, b).

The precise definitions of the two contact measures are given in Section (2). The algorithm for the docking problem will be sketched in Section (3). In Section (4) the results for some "real world" docking examples will be presented. We tested our docking algorithm with a test set that contains the test examples of Norel et al. [NLWN94] and Fischer et al. [FLWN95]. In 32 of 35 examples the best conformation with respect to the fitness function was an approximation of the real conformation. All experiments were carried out with one fixed parameter set. The parameter set was optimized so that excellent results are attained for examples where molecule B is of small or medium size ($\|B\| < 900$ non-hydrogen atoms). In all the 25 such examples in our test set the best conformation with respect to the fitness function was an approximation of the real conformation. The rationale behind choosing the parameter set to perform well on such examples is the following. We plan to use the docking program for database screening; i.e., searching structure databases for possible small or medium sized docking partners for a given protein. The implementation of our algorithm is able to handle a list $L(B)$ of docking partners for molecule A. Further a graphical user interface for marking parts of the surface of A was implemented. If a part of the surface of A is marked, then the algorithm carries out a local docking search testing only the marked area; i.e., the user can, for example, mark the active site of an enzyme of a virus and search for possible inhibitors in databases. In Section (5) we summarize our experience with the new algorithm and discuss approaches for refining the model and some future research directions.

2 The New Complementarity Measures

In this section the new complementarity measures $GeoFit$ and $ChemFit$ are presented.

The geometric rigid body docking problem is defined as follows: Given two proteins A and B with n and m atoms, determine all transformations (rigid motions) of B with A fixed so that there is a large fit between the surface of A and the surface of B and almost no penetration of B into the interior of A. The parts of surfaces that match for a particular conformation are called *common surface* or *contact surface* of the conformation.

We now define the geometric fitness function $GeoFit$ that "measures" the size of the common surface of a given conformation $conf(A, B)$:

$$GeoFit(conf(A,B)) = \sum_{\substack{a \in A, b \in B \\ 2.75 \leq d(a,b) \leq 4.0}} C_{vdw} - \sum_{\substack{a \in A, b \in B \\ d(a,b) < 2.75}} C_{pen}.$$

For each atom pair $(a, b), a \in A, b \in B$, whose Euclidian distance $d(a, b)$ is larger than 2.75 Å and smaller than 4.0 Å, a constant C_{vdw} is added to the fitness function. Thus, the first sum in the fitness function counts the number of atom pairs that have a "van der Waals contact". The second sum represents a negative score for "overlapping" atom pairs. We presently do not take into account that atoms have different van der Waals radii, but we could easily refine our fitness function with a modest increase of running time and space requirements.

The chemical fitness function $ChemFit$ is also based on van der Waals contacts. But instead of adding a constant for an atom pair (a, b) with a van der Waals contact, a weight $ChemWeight(a, b)$ that depends on the atom pair (a, b) is added to the fitness function:

$$ChemFit(conf(A,B)) := \sum_{\substack{a \in A, b \in B \\ 2.75 \leq d(a,b) \leq 4.0}} ChemWeight(a, b).$$

The computation of the weights is based on the following classification of atoms. We assume for simplicity that all molecules are made of a set of base fragments (the amino acids, the nucleic acids, NADH, FAD, Heme and so on). Each base fragment has a fragment index $frag\_index$. The atoms of each base fragment are enumerated so that each atom has a unique "$atom\_index$". The type of an atom a of a molecule is a two dimensional vector $type(a) := (frag\_index(a), atom\_index(a))$ that contains the index "$frag\_index(a)$" of the fragment to which atom a belongs and the atom index "$atom\_index(a)$" of a.

We compute the weights $ChemWeight$ as follows: We select a set of reliable docking examples out of the Brookhaven Protein Database (PDB). For each pair $(type_i, type_j)$ of atom types the number of van der Waals contacts $no\_of\_cont(type_i, type_j)$ in the test set is determined. The number of occurences $no\_of\_occ(type_i)$ of each atom type is computed and stored in a table. The weight

$ChemWeight(a,b)$ is defined as:

$$\begin{aligned} ChemWeight(a,b) &:= ChemWeight(type(a), type(b)) \\ &:= \frac{(no\_of\_all\_occ)^2}{no\_of\_occ(type(a)) \cdot no\_of\_occ(type(b))} \\ &\quad \cdot \frac{no\_of\_cont(type(a), type(b))}{no\_of\_all\_cont}. \end{aligned}$$

Here, "$no\_of\_all\_occ$" is the number of atoms in the test set and
"$no\_of\_all\_cont$" is the number of van der Waals contacts in the test set. Out of several tested statistical weight measures the above measure yields the best docking results.

3 The Algorithm

First, we describe the data structures built in a preprocessing step for the geometric and chemical fitness test. Secondly, we outline the technique for selecting a discrete set of conformations to be tested. Finally we sketch a parallel version of our docking algorithm.

The geometric fitness test: For a point p its *contact value* is defined as

$$ConVal(p) := \sum_{\substack{a \in A \\ 2.75 \leq d(a,p) \leq 4.0}} C_{vdw} - \sum_{\substack{a \in A \\ d(a,p) < 2.75}} C_{pen},$$

i.e, the contact value of p is simply the value of our geometric fitness function for a molecule consisting of a single atom which is placed at point p of the three-dimensional space.

We describe two data structures that allow to efficiently determine an approximation of the contact value $ConVal(p)$ for any p. The second data structure is faster than the first, but uses more space. For both data structures a $3D$ grid that contains molecule A is computed. The boxes of the grid have a side length of 4 Å. If all points in a box have the same contact value, we store the contact value with the box. Otherwise we store the value "Undefined" and a pointer to a local data structure for this box. The two data structures for the geometric fitness test differ in the local data structure that is added to boxes with value "Undefined." In the first data structure this local data structure is a simplified octree [FVF+90] with a constant number (default:4) of hierarchy levels. The leaves of the octree store the maximum of the contact values of the eight corners of the corresponding cube. The second data structure has a $3D$ grid (array of contact values) as the local data structure. The approximation of the contact value that is stored for a cell of the grid is the maximum of the contact values of its eight corners. It enables faster tests, but requires more storage.

In order to compute the geometric fitness of a given conformation $conf(A,B)$ the following operations have to be carried out for each atom b of B that "belongs" to the Connolly surface of B: Determine the box of the grid that contains the atom b. If the value of the box is not "Undefined", then we add this value to the fitness function. Otherwise we search the local data structure of the box for a smaller box that contains the atom and has a defined contact value. This value is added to the fitness value. The sum of all contact values is the geometric fitness value of the conformation.

Instead of considering all atoms of B, we compute only the contact values of the atoms of B that belong to the Connolly surface of B [Con1, Con2]. These atoms can be easily computed in a preprocessing step. The rationale behind looking only at atoms in the Connolly surface is that

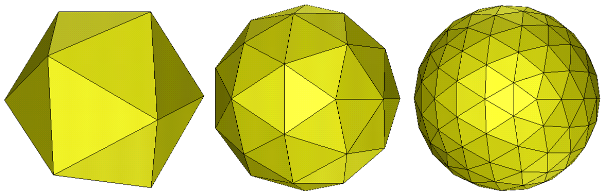

Figure 1: An icosahedron and the first two recursive refinements.

atoms of B that do not belong to the Connolly surface of B have a contact value 0 in most feasible conformations.

The chemical fitness test: The data structure for the chemical fitness test consists of two elementary data structures. The first elementary data structure is an array that contains the weights $ChemWeight(a,b)$. The second data structure is a $3D$ grid with a grid box length of 4 Å. The atoms of molecule A that are contained in the box as well as a list of pointers to the non-empty neighbor boxes, are stored in each box of the grid. In order to compute the chemical fitness of a conformation the following test has to be carried out for each atom b belonging to the Connolly surface of B: compute the box of the grid that contains b; determine all atoms stored in this or in neighbor boxes that have a van der Waals contact with b; for each atom a with this property compute the weight $ChemWeight(a,b)$ and add it to the chemical fitness function of the conformation.

How can we select the possible docking conformations? Now we describe the method for selecting the conformations $CONF$ that will be tested.

In a first step, we compute a point set P above the surface of molecule A which marks possible positions for atoms of B in the following way: We compute an almost uniformly distributed point set on the surface of a sphere s. We can determine such a point set by recursively refining an icosahedron (see Figure 1). For our purposes we take a sphere with a radius of 3.5 Å.

For each atom a of A we carry out the following test: We move the center of the sphere s to the center of atom a. For each point of the discrete surface point set of sphere s the algorithm checks if the point belongs to the so called *probe center surface*. A point belongs to this surface if the smallest distance to any atom in $A - a$ is greater or equal to 3.5 Å. We store all the points that belong to the probe center surface in a list L. For each point p in the list L the contact value $ConVal(p)$ is computed. We select the points with "large" contact values (default: ≥ 12) and store them in the point set P (see Figure 2). The points that have such large contact values are usually located in invaginations of the surface of A.

In a second step the algorithm "matches" triples of points in P and triples of atom centers of molecule B using geometric hashing [LW88]: We compute all triangles between points of P, whose side lengths are larger than a lower bound l_l and smaller than an upper bound l_u, and store them in a hash table H. Then we do the same for the centers of the atoms of B that belong to the Connolly surface of molecule B; i.e., we compute all triangles that fulfill the above length conditions. For each of the triangles between atom centers of B, we determine all "similar" triangles in the hash table

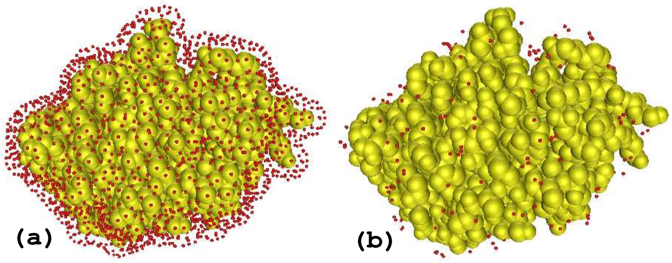

Figure 2: (a) All points on the probe sphere surface. (b) The points with contact value greater or equal to 12.

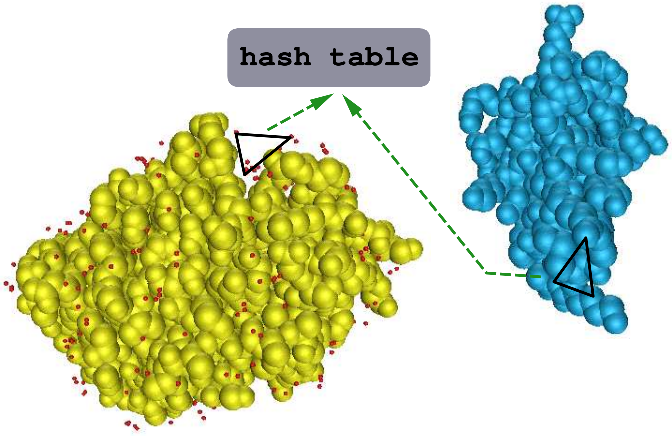

Figure 3: How to determine the transformation test set.

H. Two triangles are "similar" if their edges have almost the same lengths. For each pair of similar triangles (t_1, t_2) a transformation is computed that maps t_1 onto t_2. Since the triangles are similar but not equal, there are different ways to map the triangles. We use the centers of gravity, the normals of the triangles and angle bisectors to determine a transformation (a point = "center of gravity" and three orthonormal vectors). Thus each pair of similar triangles yields a transformation that has to be applied to molecule B. The resulting conformation is added to the test list $CONF$.

Fitness filters: In our docking algorithm we use "geometry" as a first filter; i.e., we compute the geometric fitness of the conformation set $CONF$ described above. We remove all conformations from the set $CONF$ whose geometric fitness values are smaller than a constant C_{geo}. For each of the remaining conformations the chemical fitness function $ChemFit(conf(A, B))$ and the weighted sum

$$\begin{aligned} Fit(conf(A,B)) &= w_{geo} \cdot GeoFit(conf(A,B)) + \\ &\quad w_{chem} \cdot ChemFit(conf(A,B)) \end{aligned}$$

are computed. The algorithm outputs the 25 conformations with the largest fitness value.

The parallel version of the docking algorithm: The docking algorithm described above can be easily parallelized, by splitting the list of geometric fitness tests. A master

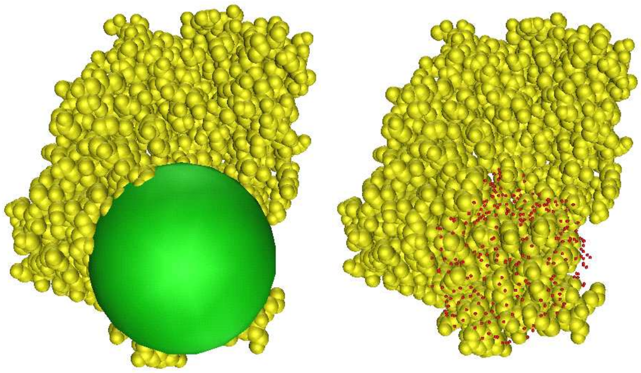

Figure 4: Using a graphical interface, we can mark parts of the surface of molecule A that should be tested.

processor distributes the work between a set of clients and coordinates the clients. Each client builds the data structures for the fitness tests in a preprocessing step. Then the master processor informs the client which part of the transformation list it should work on by sending it an integer i. This integer is the list number where the client should start. The client stops at $i + STEP$, where $STEP$ is a small integer. The client informs the master that it has finished by returning an integer. Either all the work has been completed — in this case the master informs the client that it should send its list of the best transformations — or there is an unprocessed part of the transformation list — then the master sends a new start number to the client. The master collects all results from the clients and computes a list of the best transformations. There is no communication between the clients. The message passing is handled by PVM routines [Sun90].

By choosing a suitable small constant $STEP$, the load of the clients is well balanced, but the communication overhead is still modest. The first version runs on a cluster of workstations with processors that have different performance values (SUN and SGI workstations). Hence, it is difficult to prove precisely how the speedup behaves, but our experience seems to imply that the speedup will be greater than 90 % for a small number of processors (< 32).

Some Important Features of the Implementation: The implementation of our algorithm is able to handle a list $L(B)$ of docking partners for molecule A; i.e., the algorithm can solve the 1-to-n docking problem and compute the "best" docking partners for A contained in the list $L(B)$ of molecules. Further a graphical interface for marking parts of the surface of A has been implemented (see Figure 3). If a part of the surface of A is marked, then the algorithm carries out a local docking search testing only the marked area. These features enable the user to search for possible "inhibitors" of enzymes; i.e., the user can mark for example the active site of an enzyme of a virus and search for possible inhibitors in databases.

Furthermore, our docking software package called "Parallel Protein Puzzle (PPP)" offers a graphical user interface

Test Set of Docking Complexes

Complex	Receptor	Atoms	Ligand	Atoms	RMSD (Å)
1abi*	Hydrolase α-Thrombin (LH)	2304	Inhib. (I)	152	2.3
1abi	Hydrolase α-Thrombin (H)	2039	Chain (L)	265	2.3
1acb	Hydrolase α-Chymotripsin (E)	1769	Eglin C (I)	522	2.0
1cho	α-Chymotrypsin (E)	1750	Turkey 2 Ovomucoid 3rd Domain (I)	400	1.8
1fdl	IG*G1 fab fragment (LH)	3308	2-Lysozyme (Y)	1001	2.5
1tec	Thermitase Eglin-c (E)	2004	Leech (I)	522	2.2
1tgs	Trypsinogen (Z)	1646	Pancreatic Secr. Trypsin Inhib. (I)	416	1.8
1tpa	Anhydro-Trypsin (E)	1628	Trypsin Inhib. (I)	454	1.9
2hfl	IG*G1 fab fragment (LH)	3227	Lysozyme (Y)	1001	2.5
2igf	IG*G1 FAB Fragment (LH)	3378	Myohemerythrin 69-87 (P)	58	2.8
2kai	Kallikrein a (AB)	1799	Bovine Pancreatic Trypsin Inhib. (I)	438	2.5
2mhb	Hemoglobin α-Chain (A)	1178	β-Chain (B)	1113	2.0
2ptc	β-Trypsin (E)	1629	Pancreatic Trypsin Inhib. (I)	454	1.9
2sec	Subtilisin Carlsberg (E)	1920	N-Acetyl Eglin c (I)	530	1.8
2sic	Subtilisin (E)	1938	Subtilisin Inhib. (I)	764	1.8
2sni	Subtilisin Novo (E)	1938	Chymotrypsin Inhib. (I)	513	2.1
2tgp	Trypsinogen (Z)	1629	Pancreatic Trypsin Inhib. (I)	454	1.9
2utg	Uteroglobin Chain (A)	548	Chain (B)	548	1.6
3apr	Acid Proteinase (E)	2403	Reduced Peptide Inhib. (I)	57	1.8
3dfr	Dihydrofolate Reductase	1294	Methotrexate	81	1.7
3hfm	IG*G1 fab Fragment (LH)	3295	Lysozyme (Y)	1001	3.0
3sgb	Serine Proteinase (E)	1310	Potato Inhib. PCI-1 (I)	380	1.8
3tpi	Trypsinogen (Z)	1629	Pancreatic Trypsin Inhib. (I)	454	1.9
4cpa	Carboxypeptidase	2437	Potato Carboxypeptidase A Inhib. (I)	289	2.5
4hvp	HIV-1 Protease Chain (A)	758	Chain (B)	758	2.3
4mbn	Metmyoglobin	1217	Heme	44	2.0
4phv*	HIV-1 Protease (AB)	1520	Inhib. (I)	92	2.1
4phv	HIV-1 Protease Chain (A)	760	Chain (B)	760	2.1
4sgb	Serine Proteinase (E)	1310	Potato Inhib. PCI-1 (I)	380	2.1
4tpi	Trypsinogen (Z)	1629	Pancreatic Trypsin Inhib. (I)	471	2.2
5hmg	Influenza Virus Hemaglutinin (E)	2532	Chain (F)	1418	3.2
6tim	Triosephosphate Isomerase Chain (A)	1883	Chain (B)	1894	2.2
8fab	IG*G1 FAB Fragment Chain (A)	1544	Chain (B)	1635	1.8
9ldt	Lactase Dehydrogenase Chain (A)	2568	Chain (B)	1624	2.0
9rsa	Ribonuclease Chain (A)	951	Chain (B)	951	1.8

Table 1: *Columns:* (1) PDB code of the molecular complex. (2) The name of the receptor A. (3) Number of atoms of A (without hydrogen atoms). (4) The name of the ligand B. (5) Number of atoms of B (without hydrogen atoms). (6) The resolution of the complex.

for starting and controlling the docking tests and for setting the program parameters.

4 Docking Examples

We now summarize our "real world" experiments. We present results on 35 docking complexes and on 11 examples of unbound ligand and receptor pairs.

Docking of receptor-ligand complexes: Our test set containing the test set of Norel et al. [NLWN95] and the test set of Fischer et al. [FLWN95] is listed in Table (1). In all 35 docking experiments the complete surface of A has been tested against the complete surface of B. All experiments were carried out with one fixed parameter set. Note that the set of docking complexes that has been used to compute the weights $ChemWeight$ does not contain our test set. In Table (2) and (3) we compare our results with the results of Norel et al. [NLWN95], with the results of Fischer et al. [FLWN95] and with the results of Meyer et al. [MWS96]. The columns in Table (2) contain the ranks of the lowest ranking solutions with root-mean-square (RMS) deviation smaller than X Å and the RMS-deviations (computed with all atoms of B) of these solutions. Norel et al. [NLWN95], Fischer et al. [FLWN95] and Meyer et al. [MWS96] used a tolerance value X of 3.0 Å for computing the ranking. The ranking of the geometric approach of Lenhof [Len95] (Column (6)) was computed with a tolerance value X of 5.0 Å. The new results (Column (7)) were computed with $X = 3.0$ Å. Note that even conformations with more than 5.0 Å may be good approximations of the real conformations. See for example β-Trypsin and the Pancreatic Trypsin Inhibitor (2ptc). The inhibitor has the shape of a long wedge. Thus even small rotations of the inhibitor may cause "large" RMS-deviations. Therefore, for larger ligands it may make more sense to measure only the RMS-deviation of the atoms near the contact surface.

For completeness the running times of the various docking algorithms are listed in Table (3). Note that the tests have been carried out on different hardware platforms. Therefore, a fair comparison of the running times is not possible. The results of Meyer et al. have not yet been published. The author received a list of the ranks and the corresponding RMS-deviations without the running times from M. Meyer.

The comparison of the pure geometric approach and the new approach with a "mixed" fitness function shows that adding the chemical component improves the ranking significantly. See for example the ranking of 2hfl in Table (2). Testing the chemical fitness is more time consuming than

Ranking and RMS-Deviation

Complex	[NLWN95]		[NLWN95]		[FLWN95]		[MWS96]		Len95		Len96	
	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS
1abi*	—	—	—	—	—	—	—	—	—	—	1	1.16
1abi	2	0.57	2	0.57	—	—	—	—	—	—	1	0.95
1acb	5	1.78	5	1.78	—	—	1	0.98	—	—	1	1.26
1cho	—	—	—	—	6	0.80	1	0.85	1	2.96	1	2.89
1fdl	2900	2.17	3455	2.17	—	—	33	2.01	—	—	126>25	1.77
1tec	6	2.36	6	2.36	134	0.69	1	1.88	1	1.25	1	0.80
1tgs	2	1.68	2	1.68	1	0.72	1	0.31	1	1.91	1	2.98
1tpa	1	0.61	1	0.61	—	—	1	0.63	—	—	1	1.43
2hfl	24	2.07	25	2.07	—	—	28	2.56	54	3.16	1	2.46
2igf	—	—	—	—	1	1.03	1	0.61	—	—	1	1.47
2kai	125	2.18	229	2.18	—	—	1	0.89	4	3.68	1	2.38
2mhb	2	0.69	2	0.69	1	0.79	1	0.68	1	2.16	1	2.03
2ptc	2	1.27	2	1.27	3	1.15	1	1.58	1	4.35	1	1.28
2sec	249	2.99	249	2.99	8	1.06	1	1.18	1	1.02	1	1.43
2sic	2	2.08	2	2.08	—	—	1	1.95	—	—	1	0.79
2sni	3	1.16	3	1.16	—	—	1	1.22	1	1.28	1	2.05
2tgp	2	0.50	2	0.50	—	—	1	0.66	1	3.65	1	2.52
2utg	—	—	—	—	—	—	1	0.57	1	2.22	1	1.60
3apr	—	—	—	—	—	—	1	0.44	1	3.04	1	2.19
3dfr	—	—	—	—	7	0.65	1	0.61	1	1.31	1	0.70
3hfm	104	0.94	106	0.94	—	—	?>75	0.82	13	1.59	1	0.90
3sgb	—	—	—	—	—	—	—	—	4	1.28	1	0.48
3tpi	—	—	—	—	—	—	—	—	1	2.26	1	1.38
4cpa	3	1.88	3	1.88	147	1.51	1	1.20	11	3.31	1	2.68
4hvp	2	1.64	2	1.64	1	0.92	1	0.63	1	1.62	1	1.19
4mbn	—	—	—	—	1	0.48	1	0.97	1	1.89	1	1.66
4phv*	—	—	—	—	1	0.75	1–	0.90	—	—	1	1.19
4phv	—	—	—	—	327	1.25	1	0.67	—	—	1	0.83
4sgb	72	3.59	72	3.59	6	1.09	1	0.51	1	4.53	1	2.28
4tpi	2	1.40	2	1.40	2	0.91	1	0.79	1	1.52	1	0.64
5hmg	2	1.37	2	1.37	—	—	—	—	—	—	?>25	—
6tim	2	0.85	2	0.85	—	—	—	—	—	—	1	1.62
8fab	4	2.00	4	2.00	—	—	—	—	—	—	1	1.92
9ldt	2	2.08	2	2.08	—	—	—	—	—	—	1	2.65
9rsa	10	1.63	10	1.63	—	—	—	—	—	—	?>25	—

Table 2: *Columns:* (1) PDB code of the molecular complex. (2) The results of the algorithm of Norel et al. [NLWN95] (with connectivity). Each column contains the ranks of the lowest ranking solutions with RMS-deviation smaller than 3.0 Å and the RMS deviations of these solutions. Column (3) shows the results of Norel et al. [NLWN95] without connectivity. Column (4) presents the results of the (geometric) docking algorithm of Fischer et al. [FLWN95]. Column (5) contains the results of the algorithm of Meyer et al. [MWS96], Column (6) the results of the pure geometric approach of Lenhof [Len95] and Column (7) the new results using the weighted fitness function Fit.

the geometric fitness test. Therefore, the running times for the mixed fitness function may sometimes be larger than the running times for the pure geometric docking approach by a factor of 10. See for example the running times of 1cho. The average running time "increase" factor is much smaller (≈ 3).

The comparison of the rankings of the six docking algorithms shows that only the new correlation approach of Meyer et al. yields results of the same quality as our mixed fitness function approach. In 32 of 35 test examples the first (most probable) conformation on the result list of our algorithm was an approximation of the real conformation. Since we plan to use the algorithm for testing lists of small and medium size (< 900 atoms) inhibitor candidates, we have chosen a parameter set for the experiments that gives excellent results when B is a small- or medium-sized molecule. All experiments have been carried out with this parameter set. Note that all examples with $\|B\| < 900$ have been solved "optimally"; i.e., the best conformation with respect to the fitness function was an approximation of the real conformation. Nevertheless the algorithm also yields good results for most "large" docking examples. But for large examples the running time and the quality of the results can be tremendously improved by changing the parameter set, especially the distance parameters for computing the "triangles". Adapting the parameters for large molecules B (one new parameter set) improves for example the ranking of 1fdl ($126 \rightarrow 17$, RMSD 1.53 Å), the ranking of 5hmg ($> 25 \rightarrow 1$, RMSD 2.41 Å) and the ranking of 9rsa ($> 25 \rightarrow 12$, RMSD 2.12 Å).

Docking of unbound receptors and ligands: The docking of receptor-ligands pairs where the spatial structure of the "docking" complexes are known are only test cases for the docking approaches. The ultimate goal of the docking research is the development and implementation of docking algorithms that are able successfully to predict docking reactions where the structures of the docking complexes are unknown. In the above 35 test examples the input for the docking algorithm was the 3D structures of A and B found

Running Times

Complex	[NLWN95]	[NLWN95]	[FLWN95]	[AHP+95]	Len95	Len96
1abi*	—	—	—		—	5.08(4)
1abi	35.30	20.06	—		—	11.01(4)
1acb	111.24	38.18	—		—	17.04(4)
1cho	43.54	15.54	10.48		3.21(2)	9.33(6)
1fdl	262.00	117.48	—		—	175.53(4)
1tec	83.42	28.42	81.54		5.00(5)	12.28(4)
1tgs	63.18	24.00	11.00		4.22(5)	4.56(4)
1tpa	60.48	23.06	—		—	13.56(4)
2hfl	304.24	135.00	—		227.56(4)	101.01(6)
2igf	—	—	51.18		—	3.53(2)
2kai	77.36	25.54	—		7.15(4)	20.08(4)
2mhb	241.24	111.30	5.42		7.07(6)	90.09(5)
2ptc	64.48	25.42	24.36		4.32(2)	6.25(4)
2sec	82.52	25.54	82.06		13.37(4)	18.04(4)
2sic	164.48	62.54	—		—	45.10(5)
2sni	86.42	31.12	—		9.46(1)	8.30(5)
2tgp	49.48	17.42	—		5.00(5)	8.11(5)
2utg	—	—	—		4.11(6)	13.26(4)
3apr	—	—	—		4.57(1)	3.38(2)
3dfr	—	—	10.48		1.23(1)	2.25(1)
3hfm	353.30	158.42	—		51.51(4)	79.56(7)
3sgb	—	—	—		6.25(4)	8.02(4)
3tpi	—	—	—		3.51(6)	20.48(3)
4cpa	50.30	15.06	22.36		5.27(4)	6.31(4)
4hvp	35.18	15.48	4.42		23.31(1)	20.22(5)
4mbn	—	—	8.30		1.32(1)	1.18(1)
4phv	—	—	5.06		—	20.31(4)
4phv	—	—	7.00		—	5.46(1)
4sgb	28.18	9.30	22.24		4.34(4)	5.07(4)
4tpi	58.18	21.24	13.18		10.28(1)	14.10(4)
5hmg	574.06	308.36	—		—	327.43(4)
6tim	539.36	295.7	—		—	199.39(7)
8fab	128.7	68.54	—		—	47.42(7)
9ldt	713.18	457.36	—		—	921.56(4)
9rsa	126.30	44.42	—		—	74.32(4)

Table 3: The running times of the different algorithms (in minutes.seconds). *Columns:* (1) PDB code of the molecular complex. (2) The running times of the algorithm of Norel et al. [NLWN95] with "connectivity". Column (3) shows the running times of the algorithm of Norel et al. [NLWN95] without "connectivity". The times have been measured on a PC clone 486 (66Mhz). Column (4) presents the running times of the geometric docking algorithm of Fischer et al. [FLWN95]. The times have been measured on an Silicon Graphics workstation. Column (5): The results of Meyer et al. [MWS96] are still not published. The author recieved the results by private communication. The result list did not contain running times. Column (6) the running times of the geometric algorithm of Lenhof [Len95] and Column (7) the running times of the new algorithm. The times have been measured on a cluster of SGI workstations (R4400,150Mhz). The numbers of processors that were used are given in brackets.

in the docking complex. The input of the "unbound" tests consists of 3D structures of A and B that have been elucidated separately.

We have tested 11 "unbound" receptors and ligands with a parameter set that is very tolerant as regards overlappings of atoms; for example, the "penalty" C_{pen} for overlapping atom pairs is small. The more tolerant parameter set causes a significant increase in running times because the number of completely tested conformations grows tremendously. Since we have only recently started testing unbound examples, we have not yet optimized the parameters. This will be done in the near future. Our first goal was to check how sensitive the fitness measures are to small local structure changes; i.e., what happens with the ranking when the two 3D structures are not as "complementary" as the structures that can be found in the crystal docking complexes. We have not yet attempted to improve the running times, which could be significantly improved by optimizing the docking parameters. In Table (4) we summarize the preliminary results for unbound examples. Details can be found in the full version of the paper.

5 Summary

We have presented a new algorithm for the rigid protein docking problem that yields good docking results for complexes. The first results for unbound examples are satisfactory and show that the new contact measures have the potential to overcome problems arising from overlappings and local structure changes. The geometry fitness function "counts" the number of van der Waals contacts. The chemistry fitness function measures the "chemical" probability of docking conformations. Adding the chemical valuation to the geometric fitness improves the results significantly. The obvious next step is to optimize the docking parameters for unbound examples. Future extensions include adding a fit-

Examples with Unbound Receptors and Ligands

Receptor	RMSD (Å)	Ligand	RMSD (Å)	Rank	RMSD (Å)	Running Time
1cho (E)	1.8	2ovo	1.5	1	3.21	240.13(4)
1tpo	1.7	2ptc (I)	1.9	1	1.27	235.41(4)
1tpo	1.7	4pti	1.5	78	0.87	128.15(4)
2apr	1.8	3apr (I)	1.8	5	1.67	3.22(4)
2ptc (E)	1.9	4pti	1.5	23	2.38	345.29(4)
2sni (E)	2.1	2ci2 (I)	2.0	17	1.80	166.23(4)
3cpa	2.0	4cpa (I)	2.5	?>25		
3hfm (LH)	3.0	1lyz	2.0	?>25		
5cha (A)	1.7	2ovo	1.5	7	3.83	381.41(4)
5cha (A)	1.7	1cho (I)	1.8	1	1.92	399.07(4)
9hvp (AB)	2.8	5hvp (I)	2.0	1	1.50	21.49(4)

Table 4: *Columns:* (1) PDB code of the receptor. (2) Resolution of the receptor. (3) PDB code of the ligand. (4) Resolution of the ligand. (5) The ranks of the lowest ranking solutions with RMS-deviation smaller than 4.0 Å and (6) the RMS deviations of these solutions. (7) The running times in minutes.seconds. The number of used processors is given in brackets.

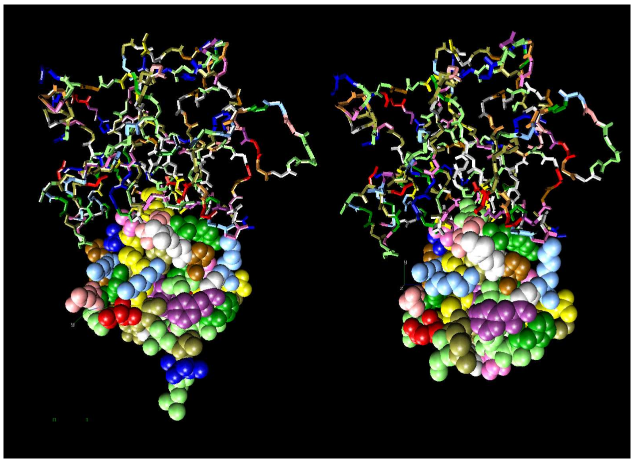

Figure 5: The snapshot on the right shows the docking complex 1CHO (A = α-Chymotrypsin and B = an inhibitor). The snapshot on the left shows the number one of our result list, when the docking is carried out with the "unbound" conformations of A (5cha) and of B (2ovo).

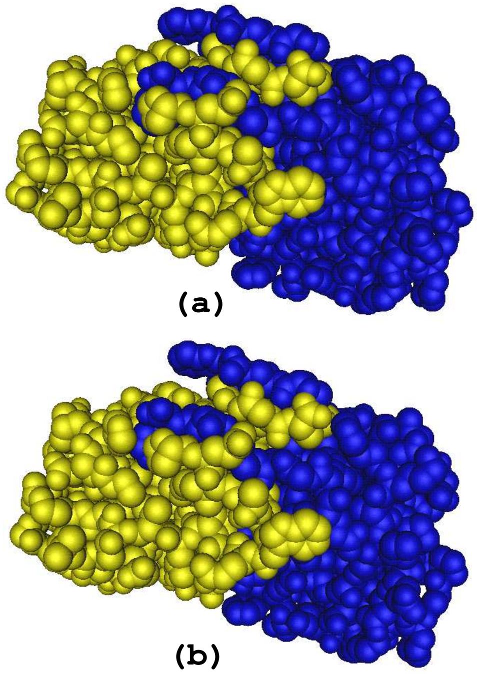

Figure 6: Example: (a) The natural docking conformation of the HIV-1 protease (dimer). (b) The best geometric fit.

ness function that judges the electrostatic interactions and implementing an energy evaluation routine for computing the final scoring of the best conformation candidates.

Acknowledgment: I thank Prof. Dr. Kurt Mehlhorn, Dr. Susanne Albers and Dipl. Chem. Oliver Kohlbacher for their comments on earlier versions of the paper. Their proposals significantly improved the readability of the paper. I am also grateful to Cand. Inform. Nicolas Boghossian for implementing the protein docking statistics.

References

[AHP+95] F. Ackermann, G. Herrmann, S. Posch and G. Sagerer: "3-D Segmentierungstechniken und vektorwertige Bewertungsfunktionen für symbolisches Protein-Protein-Docking.", in "Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism." Edited by D. Schomburg and U. Lessel, GBF Monographs, vol. 18, 1995, pp. 105–124.

[BH84] W. Bennet and R. Haber: "Structural and Functional Aspects of Domain Motion in Proteins.", CRC Crit. Rev. Biochem., vol. 15, 1984, p. 291.

[BM92] D.J. Bacon and J. Moult: "Docking by Least-Squares Fitting of Molecular Surface Patterns.", J. Mol. Biol., vol. 225, 1992, pp. 849–858.

[Con1] M.L. Connolly: "Analytical Molecular Surface Calculation.", J. Appl. Cryst., vol. 16, 1983, pp. 548–558.

[Con2] M.L. Connolly: "Solvent Accessible Surface of Proteins and Nucleic Acid.", Science, vol. 221, 1983, pp. 709–713.

[Con3] M.L. Connolly: "Shape Complementary at the Hemoglobin $\alpha_1\beta_1$ Subunit Interface.", Biopolymers, vol. 25, 1986, pp. 1229–1247.

[CDJ91] J. Cherfils, S. Duquerroy and J. Janin: "Protein-Protein Recognition Analyzed by Docking Simulations.", Proteins: Struc. Func. Genet., vol. 11, 1991, pp. 271–280.

[DSD+86] R. DesJarlais, R. Sheridan, J. Dixon, I. Kuntz and R. Venkataraghavan: "Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape.", J. Med. Chem., vol. 29, 1986, pp. 2149–2153.

[DNS+92] M. Dixon, N. Nicholsen, L. Sherwchuk, W. Baase and B. Matthews: "Structure of a hinge-bending Bacteriophage T4 Lysozyme.", J. Mol. Biol. vol. 227, 1992, pp. 917–933.

[EKS+95] M. Ester, H.P. Kriegel, T. Seidl and X. Xu: "Formbasierte Suche nach komplementären 3D-Oberflächen in einer Protein-Datenbank.", in Datenbanksysteme in Büro, Technik und Wissenschaft, GI-Fachtagung 1995, Georg Lausen (Hrsg.), Springer Verlag, pp. 373–382.

[FNN+93] D. Fischer, R. Norel, R. Nussinov and H.J. Wolfson: "3-D Docking of Protein Molecules.", Combinatorial Pattern Matching, 1993, LNCS 684, Springer Verlag, pp. 20–43.

[FLWN95] D. Fischer, S. Liang Lin, H.L. Wolfson and R. Nussinov: "A Geometry-based Suite of Molecular Docking Processes.", J. Mol. Biol., vol. 248, 1995, pp. 459–477.

[FM90] H. Faber and B. Matthews: "A Mutant T_4 Lysozyme Displays Five Different Crystal Conformations.", Nature, vol. 348, 1990, pp. 263–266.

[FVF+90] J. Foley, A. van Dam, S. Feiner and J. Hughes: "Computer Graphics: Principles and Practice.", Addison Wesley, 1990.

[GC91] M. Gerstein and C. Chothia: "Analysis of Protein Loop Closure. Two Types of Hinges Produce One Motion in Lactate Dehydrogenase.", J. Mol. Biol., vol. 220, 1991, pp. 133–149.

[HCT94] M. Helmer-Citterich and A. Tramontano: "PUZZLE: A New Method for Automated Protein Docking Based on Surface Shape Complementarity.", J. Mol. Biol., vol. 235, 1994, pp. 1021–1031.

[JK91] F. Jiang and S.-H. Kim: " Soft Docking: Matching of Molecular Surface Cubes.", J. Mol. Biol., vol. 219, 1991, pp. 79–102.

[KSE+92] E. Katchalski-Katzir, I. Sharir, M. Eisenstein, A.A. Friesem, C. Aflalo and I.A. Vakser: "Molecular Surface Recognition: Determination of Geometric Fit between Protein and their Ligands by Correlation Techniques.", Proc. Natl. Acad. Sci. USA, 1992, vol. 89, pp. 2195–2199.

[KBO+82] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge and T.E. Ferrin: "A Geometric Approach to Macromolecule-Ligand Interactions." J. Mol. Biol., vol. 161, 1982, pp. 269–288.

[KCF84] F.S. Kuhl, G.M. Crippen and D.K. Friesen: "A Combinatorial Algorithm for Calculating Ligand Binding.", J. Comp. Chem., vol. 5 (1), 1984, pp. 24–34.

[Len95] H.-P. Lenhof: "An Algorithm for the Protein Docking Problem.", in "Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism." Edited by D. Schomburg and U. Lessel, GBF Monographs, vol. 18, 1995, pp. 125–139.

[Lew91] R.A. Lewis: "Clefts and Binding Sites in Protein Receptors.", in Methods in Enzymology, Editor: J.J. Langone, vol. 202, 1991, pp. 126–156.

[LK92] A. Leach and I. Kuntz: "Conformational Analysis of Flexible Ligands in Macromolecular Receptor Site.", J. Comp. Chem., vol. 13, 1992, pp. 730–748.

[LNF+94] S.L. Lin, R. Nussinov, D. Fischer and H.J. Wolfson: "Molecular Surface Representations by Sparse Critical Points.", PROTEINS: Structure, Function, and Genetics, vol. 18, 1994, pp. 94–101.

[LW88] Y. Lamdan and H.J. Wolfson: "Geometric Hashing: A General and Efficient Model-Based Recognition Scheme.", In Proceedings of the IEEE Int. Conf. on Computer Vision, 1988, pp. 238–249.

[MWS96] M. Meyer, P. Wilson and D. Schomburg: "Hydrogen Bonding and Molecular Surface Shape Complementarity as a Basis for Protein Docking.", to appear in J. of Molecular Biology.

[MSS+89] M. Miller, J. Schneider, B. Sathyanarayana, M. Toth, G.R. Marshall, L. Clawson, L. Selk, S. Kent and A. Wlodawer: "Structure of Complex HIV-1 Protease with a Substrate-Based Inhibitor at 2.3 Å Resolution.", Science, vol. 246, 1989, pp. 1149–1152.

[NLWN94] R. Norel, S.L. Lin, H.J. Wolfson and R. Nussinov: "Shape Compementarity at Protein-Protein Interfaces.", Biopolymers, vol. 34, 1994, pp. 933–940.

[NLWN95] R. Norel, S.L. Lin, H.J. Wolfson and R. Nussinov: "Molecular Surface Complementarity at Protein-Protein Interfaces: The Critical Role Played by Surface Normals at Well Placed, Sparse, Points in Docking.", J. Mol. Biol. vol. 252, 1995, pp. 263–273.

[RSW92] J. Rini, U. Schulze-Gahmen and I. Wilson: "Structural Evidence for Induced Fit as a Mechanism for Antigen-Antibody Recognition.", Science, vol. 255, 1992, pp. 959–965.

[Sun90] V.S. Sunderam: "PVM: A Framework for Parallel Distributed Computing.", Concurrency: Practice & Experiment, vol. 2, 1990, pp. 315–339.

[SK91] B.K. Shoichet and I.D. Kuntz: "Protein Docking and Complementarity.", J. Mol. Biol., vol. 221, 1991, pp. 79–102.

[SNW95] B. Sandak, R. Nussinov and H.J. Wolfson: "An Automated Computer Vision and Robotics-based technique for 3-D Flexible Biomolecular Docking and Matching.", CABIOS, vol. 11, no. 1, 1995, pp. 87–99.

[SVC+92] A. Sali, B. Veerapandian, J.B. Cooper, D. Moss, T. Hofmann and T. Blundell: "Domain Flexibility in Aspartic Proteinases.", Proteins: Struc., Func., Genet., vol. 12, 1992, pp. 158–170.

[Wan91] H. Wang: "A Grid-Search Molecular Accessible Algorithm for Solving the Protein Docking Problem.", J. Comp. Chem., vol. 12, 1991, pp. 746–750.

[WWF+91] C. Weber, G. Wilder, B. von Freyberg, R. Traber, W. Braun, H. Widmer and K. Wuthrich: "The NMR Structure of cyclosporin. A Bound to Cyclophilin in Aqueous Solution.", Biochemistry, vol. 30, 1991, pp. 6563–6574.

[WS92] P.H. Walls and J. Sternberg: "New Algorithm to Model Protein-Protein Recognition Based on Surface Complementarity, Applications to Antibody-Antigen Docking.", J. Mol. Biol., vol. 228, pp. 227–297.