J Comput Aided Mol Des (2008) 22:815–829
DOI 10.1007/s10822-008-9216-5

A new peptide docking strategy using a mean field technique with mutually orthogonal Latin square sampling

P. Arun Prasad · N. Gautham

Received: 14 June 2007 / Accepted: 15 April 2008 / Published online: 9 May 2008

Abstract The theoretical prediction of the association of a flexible ligand with a protein receptor requires efficient sampling of the conformational space of the ligand. Several docking methodologies are currently available. We propose a new docking technique that performs well at low computational cost. The method uses mutually orthogonal Latin squares to efficiently sample the docking space. A variant of the mean field technique is used to analyze this sample to arrive at the optimum. The method has been previously applied to explore the conformational space of peptides and identify structures with low values for the potential energy. Here we extend this method to simultaneously identify both the low energy conformation as well as a 'high-scoring' docking mode. Application of the method to 56 protein–peptide complexes, in which the length of the peptide ligand ranges from three to seven residues, and comparisons with Autodock 3.05, showed that the method works well.

Keywords Peptide docking · Mean field technique · MOLS sampling · Computational drug design

Introduction

Over the last few decades, a variety of tools have been developed to computationally address the problem of docking a ligand, such as a drug, in the receptor site, for example the active site in a protein [1, 2]. This docking problem has two facets. The first is to find an energy function or a scoring function for which the global minimum corresponds to the experimentally observed structure of the ligand–receptor complex. The second is to devise an algorithm to find the global minimum of this energy function. If we assume a rigid ligand with known structure, and likewise a known rigid protein structure, for a given energy function the variables to be optimized are just the three translational vectors to position the ligand in the active site, and the rotational axis and angle for the relative orientations between two molecules. However, ligand molecules, for example peptides, are usually small and flexible, and tend to adopt different structures in the vicinity of the receptor site. This adds complexity to the docking protocol, which must now discriminate between the different structures as well as different binding modes.

Here, we present a novel 'rigid receptor flexible ligand' docking method that uses mutually orthogonal Latin squares (MOLS) to simultaneously sample both the 'docking space' and the conformational space of the ligand. The sample so obtained is analyzed using a variant of the mean field technique [3], and the globally optimum conformation as well as docking pose of the ligand in the receptor site is identified. The method is an extension of the MOLS technique developed in our laboratory to explore the conformational space of peptides and small proteins [4]. In this paper we develop it further to address the docking problem, specifically for the docking of oligopeptides in proteins. In the next section we describe the method. We then describe its application to 56 test cases selected from Protein Data Bank (PDB) [5]. Finally we compare the performance of the present algorithm to that of Autodock 3.05.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9216-5) contains supplementary material, which is available to authorized users.

P. Arun Prasad · N. Gautham (✉)
Centre of Advanced Study in Crystallography and Biophysics, University of Madras, Guindy Campus, Chennai 600025, India
e-mail: n_gautham@hotmail.com

Methods

The MOLS method

The MOLS method has been presented in detail elsewhere [4, 6, 7]. For completeness the description is repeated in the supplementary material. Here we give the following summary. Given a conformational search space of N^M points and a scoring function, the MOLS algorithm calculates the value of the scoring function at about N^2 points, and analyses these to obtain the conformation corresponding to the minimum of the function. In the case of a function with multiple minima, as in the case of the conformational energy function, this set of calculations may be repeated several times to completely explore the search space and identify all the minima. The method is very fast. For example, the method was applied to the tripeptide $(Ala)_3$. A 'brute force' exhaustive search of the complete conformational space of this molecule took about 800 min of computation, while the MOLS method took only 19 min [4]. Both methods identified all the minima in the conformational space. Further evidence that the method is exhaustive was obtained when the method was applied to explore the conformational space of several oligopeptides, including Met-enkephalin and Leu-enkephalin [6]. In all cases the method identified the experimental structures and the structures predicted by other computations. In addition, the MOLS calculations identified several other low energy and physically meaningful structures [6].

In extending this method to the docking problem, two changes are made. Firstly, the search space is now expanded to include the 'docking space'. If the conformation of the peptide is specified by the M torsion angles θ_1 to θ_M, six additional parameters describe its pose in the receptor site, three for the position and three for its orientation, making a total of $M + 6$ dimensions in the search space (θ_1 to θ_{M+6}). If each dimension is sampled at N intervals, the volume of the search space is $(N)^{M+6}$. The MOLS technique calculates the value of the scoring function at $(N)^2$ points in this space, and analyses them using a variant of the mean field technique, to simultaneously identify the optimum conformation of the peptide as well as its pose. The binding site can in principle be predefined using experimental data, e.g., site-directed mutagenesis, chemical cross-linking, protein family comparisons, or through computational predictions of the binding sites [8]. In the present instance, i.e. in the test cases reported below, the binding site is derived from the coordinates of the native ligand in the published structure of the complex, and the search space is defined by a cubic box of 5 Å units centered on the centroid of the native ligand. The rotational and translational parameters inside the box and the ligand torsion angles are the variable parameters (i.e. the dimensions) in the search space. The range for each torsion angle is set to 0–360°, with a search step size of 10°. The three translation parameters along the x, y and z axes each have a range of 2.5 Å on either side of the center, and the centroid of the ligand is moved in steps of 0.14 Å. While calculating the interaction energies, the atoms that are away from box by the sum of half the length of the extended conformation of the native ligand and the maximum range of the selected interaction energy function were considered to be included in the search space. Three parameters are used to represent ligand rotation about an axis inside the cubic box, of which two represent the position of the rotation axis, and the remaining one is the angle of rotation about this axis. To define the position of the axis, an imaginary unit sphere is constructed around the center of the cubic box. The axis is formed by drawing a line from a point on the surface of the sphere to its center. The two spherical coordinates specifying the point on the surface are then used to represent the position of the axis of rotation. The range of the polar angle is from 0 to π. The range of the azimuthal angle is 0 to 2π. The range of the angle of rotation about the axis is 0 to 2π. All three angles specifying the orientation of the ligand are sampled in steps of 10°.

A second change made in the MOLS algorithm in adapting it to the docking problem is in the scoring function. Many energy functions used in docking algorithms are composed of two terms namely, the intra-molecular ligand energy and the inter-molecular interaction energy between the ligand and the receptor [9, 10]. In the present application, since the ligands belong to the class of peptides, the ECEPP/3 force field [11] is used to calculate the intra-molecular ligand energy. For the inter-molecular interaction energy the PLP scoring function [10] is used. The PLP function is selected based on the comparative evaluation of the eleven scoring functions for molecular docking [12]. The total potential energy is a weighted sum (see Results and discussion) of these two terms. In all the present docking calculations, the ECEPP energies and the PLP energies are of the same order of magnitude, ranging from −40 to 7,500 kcal/mole for the former, and −530 to 5,000 for the latter. The ECEPP/3 energy [11] is reported in units of kcal/mole, while PLP force field is reported in dimensionless units [10]. In this paper the weighted sums of the two are also reported in dimensionless units.

Since the search space is defined on a discrete grid, in each cycle of calculations the method identifies an optimum point on this grid. However the actual optimum may lie close to but not actually on the grid. The final step in identifying the optimum is therefore to perform a gradient minimization to find the nearest off-grid optimum.

Due to the use of a relatively small sample of the extremely large and complex search space, the optimum identified above may not be the global minimum, but only one of the numerous local minima. To identify another, the calculations may be repeated with a different set of

mutually orthogonal Latin squares. We have shown for peptide conformations [6] that repeating the calculations 1,500 times is sufficient to identify all the optimum points in the search space. Any further repetitions only lead to one of the points already identified. Further, not all the 1,500 solutions are unique, and many closely similar solutions occurred more than once in the set. A clustering algorithm is be used on the 1,500 solutions to identify the unique, mutually dissimilar ones.

All docking simulations were run on a Pentium 1.8 GHz processor under the Linux operating system. The test cases (Table 1) used in this study were selected from the PDB using the following criteria. Only structures of protein–peptide complexes with resolution better than 2.25 Å were considered, and all instances of unusual and modeled peptide ligands were avoided. The peptide lengths range from three to seven residues. Since the proposed docking protocol falls in the category 'rigid receptor flexible ligand', the protein molecule was held fixed. Atoms in the receptor site having multiple occupancies were dealt with by selecting the one that has the highest occupancy. Several reports have emphasized the importance of water molecules in the receptor site [13–15]. Therefore all water molecules in the receptor site that exhibited high occupancy and low temperature factor were retained, and considered part of the rigid receptor. The CPU time required to generate 1,500 structures was about 20 h, for the largest test cases. It may be noted that the gradient minimization procedure consumes the largest part (70–80%) of computational time.

Results and discussions

In the following discussions we specifically identify two structures out of the 1,500 predictions for each complex. One is the best sampled structure, i.e. the prediction that has the lowest all-atom RMSD with respect to the native structure. The predicted structure and pose of the ligand and protein is superposed on the native structure to obtain the best overall RMSD. The RMSD is then calculated for the ligand structure alone without any further rotation or translation. The other structure that we identify in the discussions is the prediction that has the lowest total energy, as defined in the 'Methods' section, of all the 1,500 predictions for each test case.

Weighting scheme

We tried different weighting protocols for the combination of intra-molecular ligand energy (ECEPP/3 force field) and inter-molecular interaction energy (PLP scoring function). Table 2 shows the series of weighting protocols tried and the respective prediction results. Chung and Subbiah [16] have suggested that structures predicted within 2.00 Å backbone RMSD are in good agreement with the respective native counterparts. According to this criterion, we picked the pair of weights 1.0 for the intra-molecular energy and 3 for the inter-molecular energy as the most suitable for the calculations, since this pair of weights yielded the maximum number of cases predicted with less than 2.00 Å backbone RMSD that also fall in the lowest energy bin. The following discussions pertain to structures obtained using this weighting scheme.

Overall results

The most important requirement of a docking calculation is its ability to distinguish the real binding conformation and pose of the ligand on the protein from nonspecific and/or energetically unfavorable ones. Ideally, the method should predict the crystal structure (or a structure with very low root mean square deviation from the crystal structure) as the one with optimum energy. In other words, the best sampled structure and the lowest energy structure should be the same. Here, this is the case in 9 of the 56 test structures (Table 3). In the other 47 cases, the method finds at least one solution that has a low energy, as well RMSD less than 4.26 Å with respect to the crystal structure. The best sampled structures, positioned and oriented as in the receptor site and superposed without rotation or translation on the native structure, for all the 56 test cases are shown in Fig. 1. As mentioned above, the method found 'exact' solutions to nine of the test cases in which the best sampled structure is the same as the lowest energy structure. These nine structures are 1B58, 1B4Z, 1B51, 1B5I, 1JET, 1SUA, 1OLC, 1TG4 and 2ER6. The corresponding all atom RMSD with the respective native structures are 0.89, 1.09, 1.04, 0.94, 0.72, 1.64, 1.31, 2.33 and 3.00 Å.

According to the criterion of Chung and Subbiah [16], the backbone RMSD of the best sampled structure is within 2.00 Å in 39 of the 56 cases, and within 2.50 Å in 48 of the 56 cases. Out of the former 39 cases, all 20 of the 20 tripeptide, 7 of the 8 tetrapeptide, 7 of the 13 pentapeptide and 5 of the 8 hexapeptide complexes have their backbone RMSD within 2.00 Å. In the case of heptapeptides, three of the seven were predicted with backbone RMSD within 2.50 Å. The backbone RMSD of the predicted lowest energy structure is within 2.00 Å in 27 of 56 cases (Table 2).

In 41 of the 56 cases, the best sampled structure is in the top 10% when ranked in terms of the energy. Figure 2 shows a contour plot of the number of structures that fall in specific bins of total docked energy and well as of specific all-atom RMSD values. The low energy docking solutions consist both of the conformations that belong to the native binding

Table 1 The 56 protein–ligand complexes used as test cases in this study, grouped by the length of the peptide ligand

PDB ID	Peptide sequences	No. of torsion angles	Proteins	Resolution (Å)
Tri-peptide complexes				
1A30	EDL	13	HIV-1 protease	2.00
8GCH	GAW	8	Complex of alpha-chymotrypsin	1.60
1B05	KCK	15	Oligopeptide binding protein (oppA)	2.00
1B32	KMK	17	Oligopeptide binding protein (oppA)	1.75
1B3F	KHK	16	Oligopeptide binding protein (oppA)	1.80
1B3G	KIK	16	Oligopeptide binding protein (oppA)	2.00
1B3L	KGK	14	Oligopeptide binding protein (oppA)	2.00
1B46	KPK	13	Oligopeptide binding protein (oppA)	1.80
1B4Z	KDK	16	Oligopeptide binding protein (oppA)	1.75
1B51	KSK	15	Oligopeptide binding protein (oppA)	1.80
1B58	KYK	16	Oligopeptide binding protein (oppA)	1.80
1B5I	KNK	16	Oligopeptide binding protein (oppA)	1.90
1B5J	KQK	17	Oligopeptide binding protein (oppA)	1.80
1B9J	KLK	16	Oligopeptide binding protein (oppA)	1.80
1JET	KAK	14	Oligopeptide binding protein (oppA)	1.20
1JEU	KEK	17	Oligopeptide binding protein (oppA)	1.25
1JEV	KWK	16	Oligopeptide binding protein (oppA)	1.30
1QKA	KRK	18	Oligopeptide binding protein (oppA)	1.80
1QKB	KVK	15	Oligopeptide binding protein (oppA)	1.80
1S2K	AIH	10	SCP-B a member of the Eqolisin family of peptidases	2.00
Tetra-peptide complexes				
1SUA	ALAL	12	Calcium-independent subtilisin BPN	2.10
1TJ9	VARS	14	Daboia russelli pulchella phospholipase A2	1.10
1TK4	AIRS	15	Daboia russelli pulchella phospholipase A2	1.10
1NRS	LDPR	15	Human alpha thrombin	2.40
1OLC	KKKA	20	Oligopeptide binding protein (oppA)	2.10
2DQK	VLLH	15	Proteinase K	1.93
2FIB	GPRP	11	Recombinant human gamma-fibrinogen carboxyl terminal	2.10
2NPH	AETF	14	HIV-1 protease	1.65
Penta-peptide complexes				
1JQ8	LAIYS	17	Daboia russelli pulchella phospholipase A2	2.00
1JQ9	FLSYK	21	Daboia russelli pulchella phospholipase A2	1.80
1TG4	FLAYK	20	Daboia russelli pulchella phospholipase A2	1.70
1TJK	FLSTK	20	Daboia russelli pulchella phospholipase A2	1.25
1SP5	YDQIL	21	HIV-1 protease	1.80
1BHX	DFEEI	22	Human alpha thrombin	2.30
1FCH	YQSKL	22	Peroxisomal targeting signal 1 receptor	2.20
1MF4	VAFRS	18	Naja naja sagittifera phospholipase A2	1.90
1NVQ	ASVSA	13	Complex structure of checkpoint kinase chk1	2.00
1NXO	AKAIA	16	Calcium-dependent protease	2.30
2D5W	ASKTK	20	Oligopeptide binding protein (oppA)	1.30
2DUJ	LLFND	20	Proteinase K	1.67
2GNS	ALVYK	19	Daboia russelli pulchella phospholipase A2	2.30
Hexa-peptide complexes				
1AWQ	HAGPIA	15	Cypa complex	1.58
1TP5	KKETWV	27	PDZ3 domain of PSD-95	1.54
1Q3P	EAQTRL	25	Shank PDZ	2.25

Table 1 continued

PDB ID	Peptide sequences	No. of torsion angles	Proteins	Resolution (Å)
1I31	FYRALM	25	Mu2 adaptin subunit (AP50) of AP2 clathrin adaptor	2.50
1OBY	TNEFYA	22	PDZ2 of syntenin	1.85
1TDV	YWAAAA	16	Daboia russelli pulchella phospholipase A2	1.70
2FOO	EPGGSR	19	Ubiquitin carboxyl-terminal hydrolase	2.20
2FOP	EKPSSS	21	Ubiquitin carboxyl-terminal hydrolase	2.10
Hepta-peptide complexes				
1DKX	NRLLLTG	27	Domain of dnak from type 1 selenomethionyl	2.00
1P7V	PAPFAAA	14	Proteinase K	1.08
1P7W	PAPFASA	15	Proteinase K	1.02
1U8I	ELDKWAN	29	HIV-1 cross neutralizing monoclonal antibody 2F5	2.00
1KY6	FSDPWGG	19	Alpha-adaptin C	2.00
2ER6	PTEFFRE	29	Endothia aspartic proteinase	2.00
2FOJ	GARAHSS	22	Ubiquitin carboxyl-terminal hydrolase	1.60

Table 2 The results of applying different weights for intra-molecular ligand energy (ECEPP/3) and inter-molecular interaction energy (PLP)

ECEPP/3	PLP	No. exact solutions	No. of predictions with RMSD <2.00 Å	No. of predictions in lowest energy cases with RMSD <2.00 Å
1	1	9	40	18
1	2	5	41	23
1	3	9	39	27
1	3.5	4	44	20
1	4	6	40	19
2	1	5	37	22
3	1	4	43	22

The RMSD is calculated with respect to the native structure as described in the text

mode, as well as solutions that are quite different. For example, the lowest energy prediction for the test case 2FOO that has a large RMSD of 12.57 Å with the native. The energy of the native complex and the lowest energy prediction are −268.0 and −225.5 respectively. A total of nine hydrogen bonds were observed between the protein and the lowest energy peptide ligand and only seven between the protein and the native peptide ligand. A total of 17 non-bonded contacts were observed between the protein and the native peptide ligand and only 14 between the protein and the lowest energy peptide ligand (Hydrogen bonds and non-bonded contacts were analyzed using HBPLUS [17]).

Thus, overall the correlation between RMSD and energy is low. In the context of protein–protein interactions, Knegtel et al. [18] have observed that this correlation decreases with decrease in the extent of the complementary surfaces of the two interacting proteins. In order to see if a similar effect was present in these calculations, we studied the relationship between the rank (as calculated from the energy value) of the best sampled structure and the solvent accessible surface area (SASA) of the binding site of the ligand. The SASA was calculated using NACCESS [19] for all the atoms found within the search space as described 'Methods' section. No such relationship appeared. Structures ranked in the top 10% had a range of SASA values, from a high of about 4,500 $Å^2$ to a low of about 500 $Å^2$.

The complementary contact surfaces between the ligand and the receptor are substantially smaller, and therefore probably less discriminating than the case of protein–protein docking. Moreover, the peptide ligands are highly flexible, adapting their surface to optimally complement the receptor pocket, and may therefore have several nearly similar solutions [8]. In other words the energy hypersurface is rugged, with many minima.

We conclude this section by reiterating that despite the large size and the extreme unevenness of the search space, the MOLS search algorithm identifies reasonable solutions at reasonable computation cost in all the test cases. Particular examples of these are discussed below. This is followed by a comparison of the performance of this method with that of Autodock version 3.05 [20].

K×K tri-peptide ligands

Of the 20 tri-peptide complexes investigated, 17 are complexes of the sequence K×K with the oligopeptide-binding protein (oppA). This protein accommodates peptides of varying length and sequence [21] though it binds with highest affinity to peptides of length two to five residues. In the structures the ligand is completely engulfed inside the protein with its side chains pointing into a large hydrated cavity.

In every one of the 17 oppA structures, the best sampled structure was found within 2.00 Å RMSD of the native

Table 3 Summary of results for the 56 test cases

PDB ID	RMSD (Å)		RDE (%)		Energy		Native energy
	BS	LE	BS	LE	BS	LE	
1A30	1.42	1.65	48.4	46.6	−243.8	−329.8	−302.7
8GCH	0.90	1.39	70.9	68.6	−309.1	−331.8	−298.3
1B05	0.94	1.26	72.6	63.2	−429.2	−430.0	−411.0
1B32	0.95	1.21	74.2	65.9	−425.2	−430.5	−476.2
1B3F	1.14	2.02	63.2	55.0	−371.7	−420.2	−516.5
1B3G	1.07	1.74	73.0	52.1	−372.5	−394.4	−461.9
1B3L	0.99	1.74	68.4	60.2	−374.3	−419.6	−375.7
1B46	0.74	0.97	79.5	75.6	−441.2	−458.1	−488.0
1B4Z	**1.09**	**1.09**	**70.5**	**70.5**	**−495.3**	**−495.3**	**−528.9**
1B51	**1.04**	**1.04**	**71.6**	**71.6**	**−454.0**	**−454.0**	**−459.8**
1B58	**0.89**	**0.89**	**77.0**	**77.0**	**−498.9**	**−498.9**	**−542.2**
1B5I	**0.94**	**0.94**	**69.5**	**69.5**	**−485.8**	**−485.8**	**−507.4**
1B5J	1.14	1.80	68.1	60.2	−434. 8	−455.6	−490.1
1B9J	1.01	1.27	72.6	63.3	−424.3	−462.7	−497.1
1JET	**0.72**	**0.72**	**84.0**	**84.0**	**−474.4**	**−474.4**	**−476.0**
1JEU	0.98	1.08	67.8	69.6	−302.8	−487.7	−531.4
1JEV	1.17	6.56	60.9	23.6	−402.7	−461.2	−586.9
1QKA	1.18	1.22	63.3	66.1	−461.2	−513.2	−547.2
1QKB	1.12	1.28	66.7	64.7	−413.3	−482.0	−481.0
1S2K	1.35	1.81	63.1	46.6	−205.3	−236.5	−202.2
1SUA	**1.64**	**1.64**	**68.2**	**68.2**	**−278.3**	**−278.3**	**−297.5**
1TJ9	1.96	6.17	41.3	7.1	−153.9	−260.1	−91.3
1TK4	1.84	7.87	40.3	5.6	−169.3	−273.7	179.1[a]
1NRS	1.50	1.98	53.7	55.6	−319.3	−383.9	−399.0
1OLC	**1.31**	**1.31**	**62.2**	**62.2**	**−442.7**	**−442.7**	**−525.2**
2DQK	3.19	3.57	18.6	19.0	−61.1	−246.5	59369.3[a]
2FIB	1.96	2.25	40.0	49.9	−172.5	−268.4	−304.0
2NPH	1.48	8.02	60.0	4.5	−254.4	−294.6	−258.3
1JQ8	1.64	6.95	54.6	4.0	−202.5	−252.6	1580.7[a]
1JQ9	2.41	4.28	47.1	32.3	−219.9	−337.0	111728.8[a]
1TG4	**2.33**	**2.33**	**38.0**	**38.0**	**−222.1**	**−222.1**	**17519.5**[a]
1TJK	3.17	8.15	23.2	7.1	−139.7	−224.1	18079.6[a]
1SP5	3.24	7.34	27.7	10.3	380.4	−371.1	−394.7
1BHX	2.61	6.71	36.3	9.3	−163.2	−299.2	−263.0
1FCH	1.73	1.75	47.2	51.0	−471.3	−511.0	−636.4
1MF4	3.09	7.88	21.8	12.1	−264.6	−306.1	859.2[a]
1NVQ	2.09	4.31	38.6	8.9	−164.6	−211.9	−189.8
1NXO	2.79	4.02	22.2	23.2	−111.2	−235.8	−133.9
2D5W	1.56	1.89	58.5	59.0	−340.1	−436.3	−501.4
2DUJ	3.64	9.32	19.7	2.9	−190.0	−277.8	718.3[a]
2GNS	3.05	3.65	28.9	16.4	−114.1	−290.9	14730.3[a]
1AWQ	2.60	3.01	31.3	39.5	−194.3	−290.0	−345.6
1TP5	2.43	12.21	38.2	3.7	−232.7	−288.0	−361.7
1Q3P	2.97	3.63	25.5	31.4	−236.9	−308.3	−308.0
1I31	2.61	5.46	31.3	13.7	−144.9	−361.3	−337.7
1OBY	2.42	11.67	25.6	4.8	−259.0	−398.7	186572.6[a]
1TDV	3.35	11.05	16.4	2.9	−104.9	−223.1	319.0[a]
2FOO	3.04	12.57	33.7	2.1	−186.3	−225.5	−268.0
2FOP	1.68	12.01	59.8	3.8	−226.3	−245.4	−329.4
1DKX	3.05	6.16	27.8	6.0	16.5	−374.4	−425.6
1P7V	2.81	6.27	19.9	9.9	−86.9	−243.4	−190.1
1P7W	3.31	10.81	16.2	7.0	−148.0	−246.0	321.7[a]
1U8I	4.26	7.50	13.2	4.2	−128.5	−284.4	−378.8
1KY6	3.25	6.35	25.1	17.8	−217.4	−280.3	−400.1
2ER6	**3.00**	**3.00**	**22.3**	**22.3**	**−455.5**	**−455.5**	**−552.3**
2FOJ	3.42	5.43	24.2	16.4	−125.8	−301.6	−332.9

The table shows the root mean square deviation (RMSD), relative displacement error (RDE) and energy for the best sampled structures (BS), lowest energy structures (LE) and energy of the native structures for all the 56 test cases. 'Exact solutions' are shown in bold

[a] The high energy of the native structure was due to the short contacts seen in the peptide ligand

structure. All the best sampled structures fall in the top 10% when the 1,500 predictions for each are ranked according to energy. There are a total of three side chain pockets identified in oppA. Each pocket is apolar close to the peptide ligand backbone, and expands to a capacious and hydrated cavity surrounded by protein side chains [22]. Two of these pockets are occupied by lysine side chains in all 17 structures. According to the predictions, the third central pocket is preferentially occupied by small hydrophobic residues, such as valine and alanine. This is indicated by Fig. 3, which shows the frequency of structures predicted at <2.00 Å RMSD from the native.

In 16 of the 17 K×K tripeptide complexes, the lowest energy structure identified is within 2.50 Å RMSD from the native. In only one case (1JEV) the RMSD of the lowest energy structures is 6.56 Å which exhibit alternate binding modes, as discussed below.

Alternate binding modes

Since the method does not converge to a single solution, but generates hundreds of low-energy possibilities, it often detects alternate solutions that have a lower energy value than the native structure. An example of this is the structure of oligopeptide binding protein (oppA) in complex with the tripeptide KWK (IJEV) [23]. Here, the lowest energy structure identified by the algorithm probably represents an alternate binding mode. The docked energies of the native, best sampled and lowest energy structures are −586.9, −402.7 and −461.2 respectively, i.e. they may be considered approximately iso-energetic [24]. The RMSD of the best sampled and the lowest energy structure with the crystal structure are 1.17 and 6.56 Å respectively. In both modes

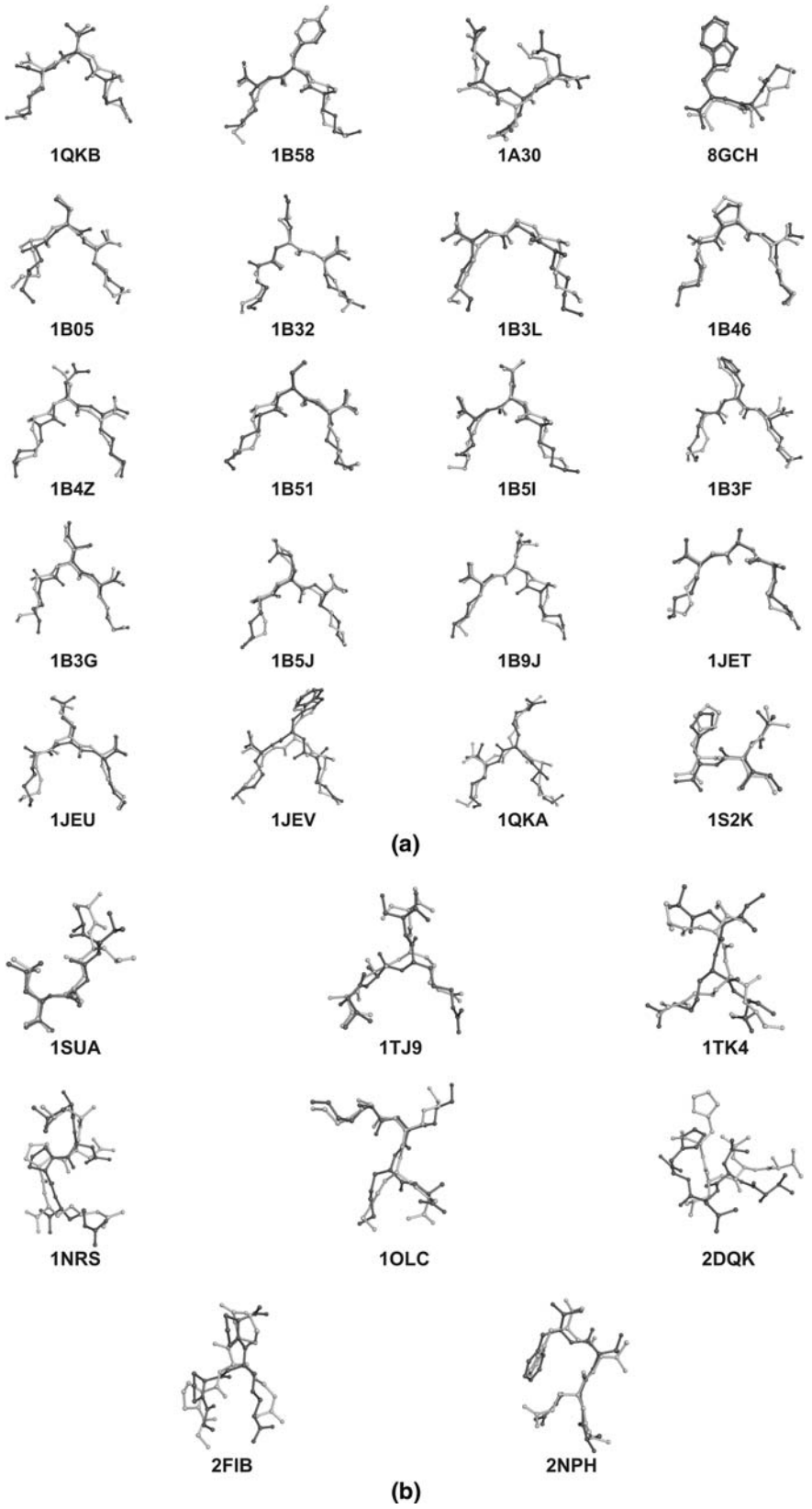

Fig. 1 Best sampled peptide ligand structures for all the 56 test cases are shown superposed without rotation and translation on the native peptide structure. The results are classified in terms of peptide length: tripeptide (**a**), tetrapeptide (**b**), pentapeptide (**c**), hexapeptide (**d**) and heptapeptide (**e**). The best sampled and the native peptides are shown in black and gray respectively

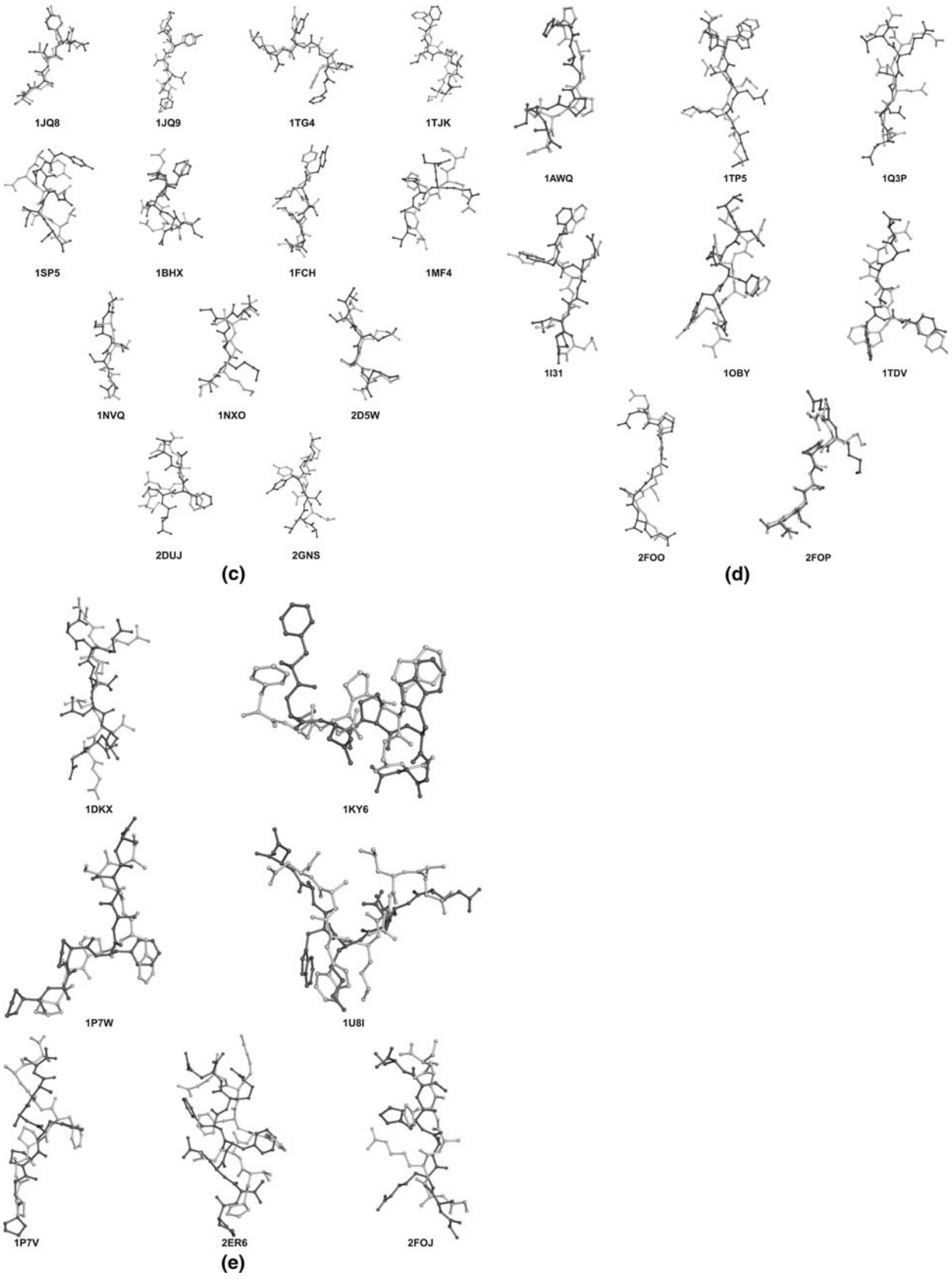

Fig. 1 continued

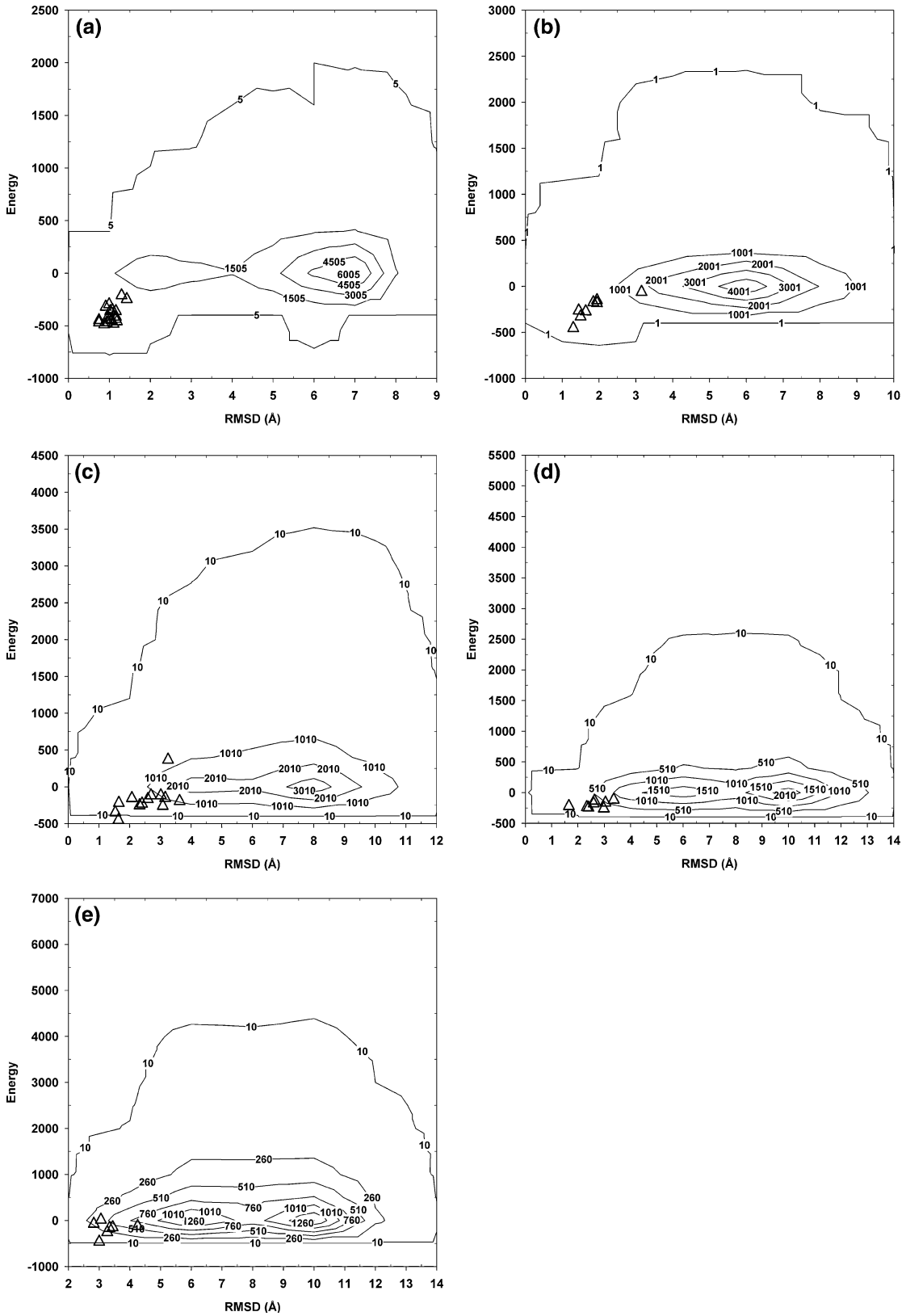

Fig. 2 Contour plot of the number of structures that fall in specific bins of total docked energy and well as of specific RMSD values for the 20 tripeptide (**a**), the 8 tetrapeptide (**b**), the 13 pentapeptide (**c**), the 8 hexapeptide (**d**) and the 7 heptapeptide (**e**) complexes. The best sampled structures are shown as open triangles

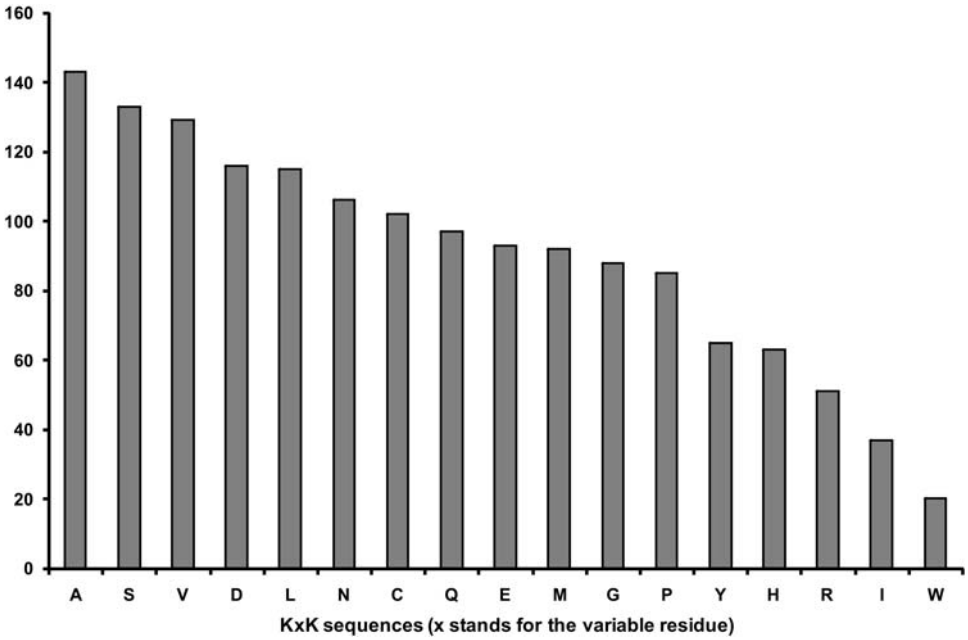

Fig. 3 Frequencies of 17 oppA complexes predicted at <2.00 Å from the native

the ligand makes the same type of hydrogen bonding patterns as seen in the native structure. However the positions of the side chains of the two residues Lys1 and Trp2 of the ligand are exactly interchanged (Fig. 4). This is surprising, in view of the completely different chemical characters of the two side chains. However, one reason for the alternate binding mode having the lowest energy may be that, in both the native and the lowest energy structure the common hydrogen bonding pattern involving residues (Gly415 and Arg413) and water molecule (W14 and W33) of the receptor protein is preserved. This common hydrogen bonding pattern is distributed through out the length of the tripeptide. Another reason could be that Trp416 of the receptor forms a total of 13 non-bonded contacts with the ligand residues Lys1 and Trp2 in the native structure. In the predicted lowest energy structure a total of eight similar non-bonded contacts are observed.

The lowest energy structures in five other cases (1SP5, 1MF4, 1OBY, 1DKX and 1P7W) also exhibit alternate binding modes. These alternate binding modes are characterized by larger values of RMSD from the native counterpart (see Table 3).

Conserved hydrogen bonding patterns

In many cases in the lowest energy structure preserves some of the hydrogen bonding patterns seen in the native structure, but not all. An example of this situation is the

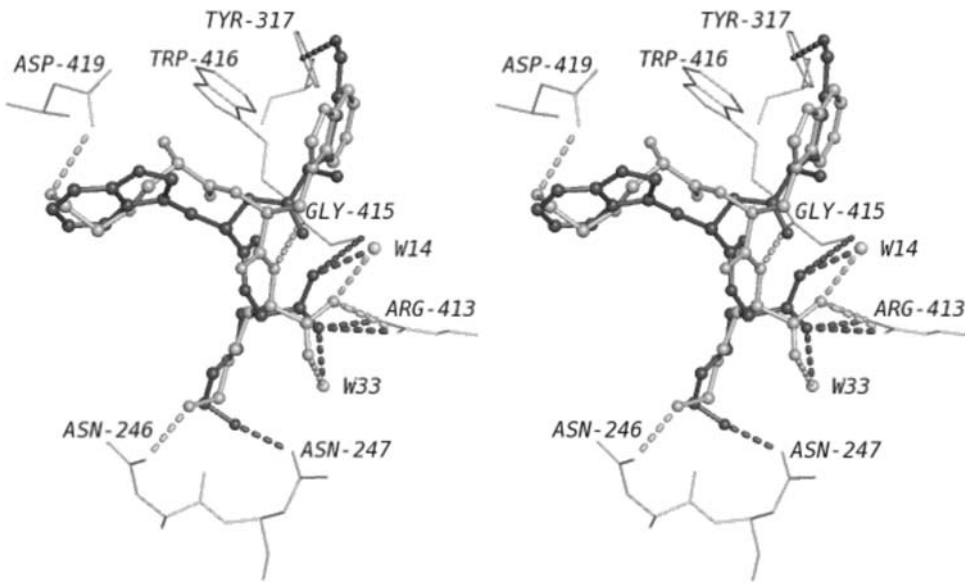

Fig. 4 Superposition of lowest energy structure of 1JEV on the respective native structures. The lowest energy peptide are shown in black and the native peptide in gray, as ball and stick model. Interacting protein residues are labeled and shown as stick model. Hydrogen bonds formed by the lowest energy peptide and native peptide are shown in black and gray respectively

structure of the test case 2FIB complexed with the tetrapeptide GPRP. Out of three main chain–main chain hydrogen bonds in the native complex, two were preserved in the lowest energy structure (Fig. 5). One of the three main chain–side chain hydrogen bonds was also preserved in the lowest energy structure. The side chain–side chain hydrogen bonds that lend specificity to the interaction are not preserved. However, in both the native and lowest energy structures, the Arg side chain of the peptide ligand forms two hydrogen bonds with the receptor residues. The docked energies of the native, best sampled and lowest energy structures are −304.0, −172.5 and −268.4 respectively. The all atom RMSD of the best sampled structure and the lowest energy structure with the native are 1.96 and 2.25 Å respectively. A different metric called the Relative Displacement Error (RDE) [25] is 49.9% for the lowest energy structure, as compared to 40.0% for the best sampled structure, indicating that the percentage of similarity with the native is better for the lowest energy structure than the best sampled structure.

A difficult target—Gly-Ala-Trp complexed with γ chymotrypsin (8GCH)

In the crystal structure of this complex [26], the position of the ligand is the average of a structure in which the ligand is covalently bound to the receptor, and one in which it is not so bonded. The COOH terminal carbon of the ligand displays short contact distances to the oxygen atom of Ser195. Docking techniques, usually developed for only non-bonded interactions, find the prediction of this complex a difficult task [14]. Despite the problematic crystal structure, the hydrophobic pocket of γ-chymotrypsin, which appears to drive the docking (Fig. 6), was identified by both the best sampled and the lowest energy structures of the MOLS predictions. The all atom RMSD of the best sampled structure with the native is 0.90 Å, while that of the lowest energy structure is 1.39 Å. The docked energies of the native, best sampled and lowest energy structures are −298.3, −309.1 and −331.8 respectively. The best sampled structure was ranked 6th in terms of energy.

Comparison with Autodock 3.05

The present MOLS algorithm was compared with Autodock version 3.05 [20]. Out of the 56 cases treated in the calculations described above, 41 were chosen for the comparison, comprising all the structures of the tetra, penta, hexa and hepta peptides complexes, but only 5 of the 20 structures from the tripeptide complexes. (17 out of 20 tripeptide complexes belong to the class of K×K sequences. Only two representatives of that class (1JEV and 1B05), plus the other three complexes that did not have a K×K ligand, were considered for the comparison.)

Both ligand and protein input files were prepared using Autodock Tool (ADT) by standard protocols described in the literature [27]. Specifically, the rotatable torsion angles were selected explicitly, and were the same as used in the MOLS method. Both the Grid Parameter File (GPF) and the Docking Parameter File (DPF) were prepared using ADT. The number of grid points in the grid box was set big enough to accommodate the extended conformation of the native ligand completely inside the grid box. The center of the grid box was set to the center of the ligand. All other grid parameter options were left at their default values. Docking was carried out using the default Genetic Algorithm (GA) parameters along with Solis and Wets local search. A total of 50 GA runs were performed. The maximum number of energy evaluation and generations were left to their default values of 250,000 and 27,000 respectively. All the docking run options were left at default values expect for the inclusion of the calculation of internal electrostatic energy.

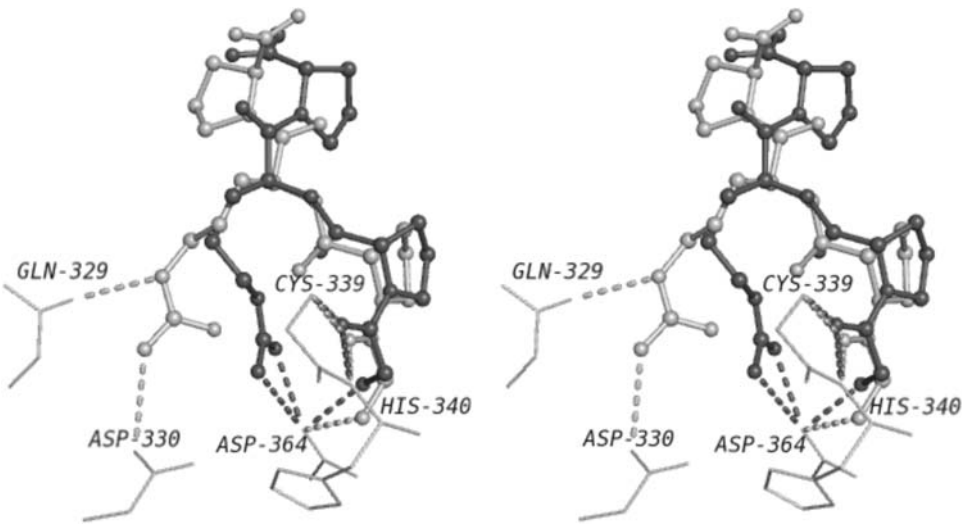

Fig. 5 Conserved hydrogen bonds: Lowest energy peptide of 2FIB superposed on the native. The three conserved hydrogen bonds between the native and the lowest energy prediction are shown. The lowest energy peptide and the native are shown in black and gray respectively

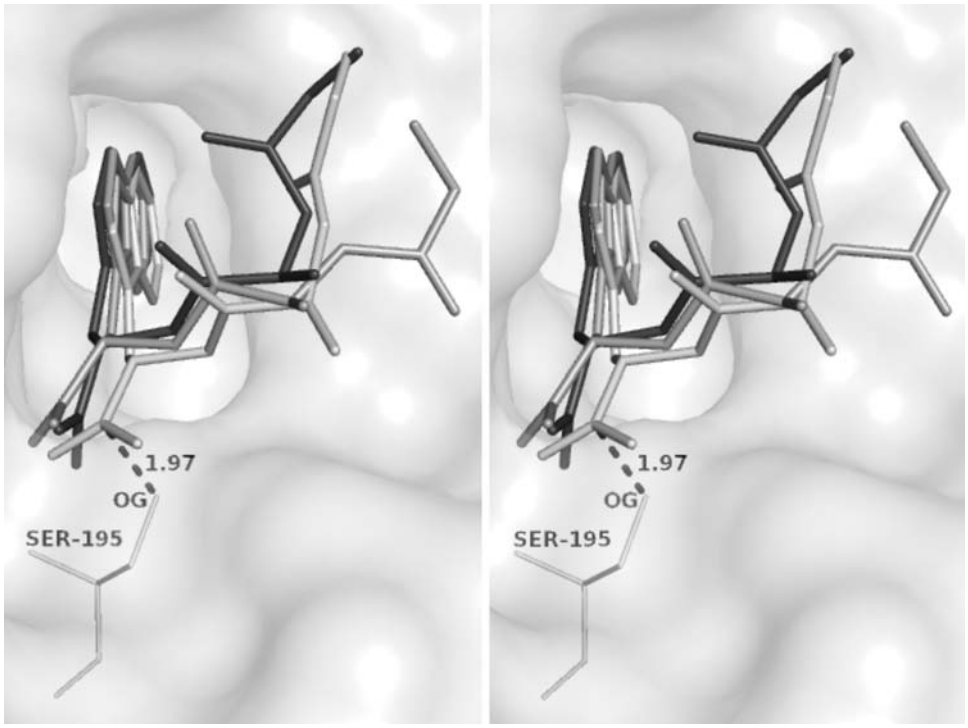

Fig. 6 Native peptide of 8GCH superposed on the best sampled structure and the predicted lowest energy structure. The native, best sampled and the lowest energy structure are shown in gray, black and dark gray respectively

Table 4 shows the results of the Autodock runs for the 41 cases, and the energy ranking of best sampled structures for both Autodock and MOLS methods. Of the 41 cases, the best sampled structure had RMSD from the crystal structure of less than 2.50 Å in 29 complexes in the case of Autodock results, and 21 complexes in the case of the MOLS results. In 10 cases, the structure best sampled by Autodock was found within the top 10% when ranked in terms of energy, whereas in MOLS method, 26 cases the best sampled structure was ranked in the top 10%. Exact solutions, i.e., solutions in which the best sampled is the same as the lowest energy structure were found by Autodock in two cases, and by MOLS in four cases. These results show that, though, in general, Autodock identifies the native binding mode slightly better than the MOLS method, the latter samples a wider range of binding modes, and ranks them better than Autodock. Autodock was able to identify alternate binding modes, as described above, in four cases (1S2K, 1SP5, 1TK4 and 1TG4) while MOLS identified six alternate binding modes (1JEV, 1SP5, 1MF4, 1OBY, 1DKX and 1P7W). In only one case the alternate mode identified by Autodock is the same as the one identified by MOLS.

Since Autodock genetic algorithm is non-deterministic, 50–100 runs were good enough to identify the solution [27, 28]. Hence for comparison studies, we fixed the maximum number of GA runs to be 50 for each test case, and the maximum number of MOLS runs was fixed to 1,500 for each test case. Figure 7 shows a comparison of the CPU time required by Autodock (50 runs) and by the MOLS (1,500 runs) to find the solution, in terms of the number of variable torsion angles considered. The average CPU time required is 0.6 h for Autodock and 11.6 h for MOLS. Thus Autodock performs 20 times faster than MOLS. However, in order to complete one single run for all the current test cases, Autodock samples 250,000 peptidyl conformations and MOLS samples only $N^2 = 1{,}369$ peptidyl conformations. Here N is the order (or size) of the MOLS square. This is set to 37 in all the test cases considered. (It should be noted that with the Autodock genetic algorithm, as with all evolutionary algorithms [20], it is a common practice to have a very high count for the number of energy evaluation ($\sim 1.5 \times 10^6$) thereby sampling more peptidyl conformations.) Thus to complete one single run, Autodock takes 42 s whereas MOLS takes only 28 s. It may also be noted that a grid based energy evaluation [29], was not used by MOLS in the current study. Using a grid based energy evaluation in docking would speed up the calculation by 180 fold [29]. Hence, with grid based energy evaluation, MOLS could find solutions with even less computation time, perhaps on par with Autodock. As the number of torsion angles increases, the CPU time also increases, though the increase is not monotonic. For an instance, the number torsion angles for the two cases 1DKX and 1Q3P is 27. However, the number of protein atoms present in the docking space for 1DKX (950) is greater than for 1Q3P (552). Thus the average MOLS docking time per cycle for 1DKX (112 s) is much greater than for 1Q3P (39 s). In another instance, an almost equal number of protein atoms are present in the docking space for the cases 1DKX (950) and 1P7V (974) having 27 and 14

Table 4 Summary of results of the Autodock runs for the selected 41 test cases

PDB ID	RMSD (Å)		Energy (kcal/mol)		Native energy (kcal/mol)	BS energy ranking (%)	
	BS	LE	BS	LE		Autodock	MOLS
1A30	0.83	1.36	1.70	−0.23	+2.76	46.00	15.13
8GCH	0.66	0.85	−13.31	−14.47	+29.70	52.00	0.40
1B05	0.84	1.25	−26.20	−28.25	−14.23	20.00	0.13
1JEV	0.82	0.95	−36.26	−40.88	−37.31	12.00	0.87
1S2K	1.16	5.45	−8.36	−11.21	−5.84	82.00	0.80
1SUA	0.99	7.49	−10.35	−13.27	−9.37	68.00	0.07
1TJ9	1.36	5.83	−19.48	−23.83	−12.23	52.00	34.33
1TK4	2.63	7.15	−14.67	−19.14	+3.05	66.00	9.33
1NRS	0.94	1.30	−17.90	−18.34	−10.46	8.00	0.73
1OLC	1.07	1.13	−42.09	−42.37	−43.05	4.00	0.07
2DQK	2.15	3.66	−8.32	−10.86	−3.46	56.00	84.47
2FIB	0.89	1.72	−18.70	−20.37	−17.00	24.00	18.93
2NPH	1.25	1.70	−6.38	−7.17	−3.81	16.00	2.27
1JQ8	1.51	2.02	5.83	−9.02	−2.09	50.00	0.40
1JQ9	1.84	4.22	−7.58	−15.04	+5.33	24.00	1.33
1TG4	2.41	10.71	−9.96	−15.97	+0.71	88.00	0.07
1TJK	3.08	11.39	−19.52	−21.16	−4.16	4.00	11.33
1SP5	1.89	8.60	−3.66	−7.36	−5.02	12.00	67.20
1BHX	2.08	3.74	+3.51	+2.98	+7.62	6.00	22.00
1FCH	**1.73**	**1.73**	**−21.89**	**−21.89**	−34.05	**2.00**	0.13
1MF4	2.14	3.76	−18.77	−19.90	+29.64	8.00	0.67
1NVQ	1.19	3.00	−5.63	−8.49	−3.88	66.00	2.73
1NXO	2.19	5.92	−9.92	−13.20	−5.54	70.00	18.00
2D5W	1.28	1.95	−70.80	−74.94	−70.31	50.00	0.53
2DUJ	6.21	8.69	1.65	−5.95	+42.09	100.00	8.60
2GNS	1.02	2.32	−14.40	−16.81	−9.31	16.00	39.40
1AWQ	0.66	3.85	−12.63	−12.71	−10.16	4.00	1.47
1TP5	3.59	8.52	−4.68	−7.81	−9.89	62.00	1.00
1Q3P	2.59	3.98	−7.86	−12.02	−7.79	34.00	2.13
1I31	3.43	5.34	−8.98	−14.40	−3.60	54.00	34.87
1OBY	7.15	10.88	−0.39	−3.92	+16.08	38.00	4.73
1TDV	4.14	7.46	−6.25	−10.13	+3.63	56.00	17.67
2FOO	1.69	2.61	−10.98	−12.83	−5.27	10.00	2.53
2FOP	1.55	2.85	−9.88	−12.48	−8.12	34.00	0.20
1DKX	2.32	4.36	−10.88	−25.06	−25.90	20.00	30.40
1P7V	4.68	8.27	−9.61	−10.82	−2.25	16.00	22.53
1P7W	3.99	4.81	−9.80	−12.54	+12.84	32.00	6.27
1U8I	4.56	9.36	−3.96	−6.47	−9.20	24.00	29.07
1KY6	1.70	2.56	−5.99	−7.85	−9.02	10.00	1.27
2ER6	**2.22**	**2.22**	**−2.90**	**−2.90**	−9.53	**2.00**	0.07
2FOJ	4.05	12.30	−11.93	−15.86	−8.81	42.00	26.53

The table shows the root mean square deviation (RMSD) for the best sampled structures (BS) and lowest energy structures (LE), energy for the (BS) and (LE), energy of the native structures and the ranking of best sampled structures of Autodock and MOLS for all the 41 test cases. 'Exact solutions' are shown in bold. The MOLS results have been abstracted from Table 3

torsion angles respectively. Due to the difference in the number of torsion angles, the average MOLS docking time per cycle for 1P7V (28 s) is much less than for 1DKX (112 s). Thus the CPU time is a function of torsion angles, number of protein atoms in the docking space and the state of the ligand before minimization.

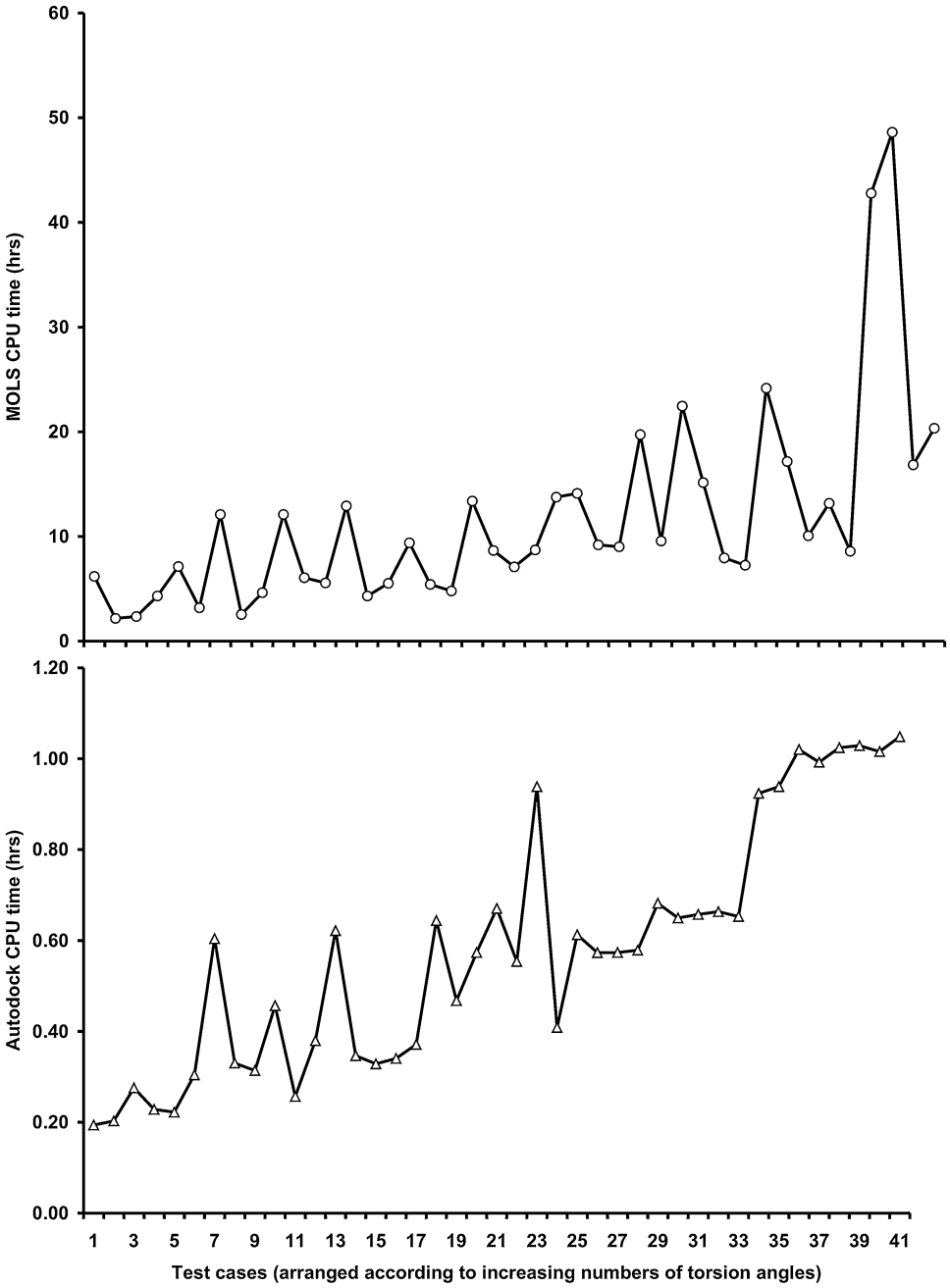

Fig. 7 The figure shows a comparison of the CPU time required by Autodock (50 runs) and by the MOLS (1,500 runs) to find a better solution, in terms of increasing number of torsion angles

Conclusion

The novel 'rigid receptor flexible ligand' docking algorithm, named MOLS, was tested by comparison with Autodock by running both on 56 protein–peptide complexes, in which the length of the peptide ligand ranges from three to seven residues. The comparison shows that the MOLS method ranks the best sampled structures better than Autodock. In terms of CPU time, Autodock finds the solutions much faster than MOLS. The MOLS method is also capable of identifying alternate binding modes. In general, it is a suitable method when it is desirable to extensively explore both conformational space and docking space simultaneously, at reasonable computational cost.

The method may be adapted for 'flexible receptor flexible ligand' docking by including the conformation of the residues lining the receptor site. However, it has been shown [30] that there is a relationship between the stability of the binding region and the changes that occur on the other parts of the protein upon binding. Hence the success of 'flexible receptor flexible ligand' algorithms may lie in understanding the loss of stability in other parts of the protein upon ligand binding. The proposed method has been demonstrated and tested for peptide ligands. It

probably may be extended to any type of ligand, with proper selection of the energy functions.

Acknowledgments We thank the Department of Biotechnology, Government of India for financial support under the grant no. BT/PR5476/BID/07/136/2004. We also thank the University Grants Commission, and the Department of Science and Technology, Government of India for support under the CAS program and the FIST program, respectively.

References

1. Taylor RD, Jewsbury PJ, Essex JW (2002) J Comput Aided Mol Des 16:151
2. Brooijmans N, Kuntz ID (2003) Annu Rev Biophys Biomol Struct 32:335
3. Koehl P, Delarue M (1994) J Mol Biol 239:249
4. Vengadesan K, Gautham N (2003) Biophys J 84:2897
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235
6. Vengadesan K, Gautham N (2004) Biopolymers 74:476
7. Vengadesan K, Gautham N (2004) Biochem Biophys Res Commun 316:731
8. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Proteins Struct Funct Genet 47:409
9. Bursulaya BD, Totrov M, Abagyan R, Brooks CL (2003) J Comput Aided Mol Des 17:755
10. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST (1995) Chem Biol 2:317
11. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA (1992) J Phys Chem 96:6472
12. Wang R, Lu Y, Wang S (2003) J Med Chem 46:2287
13. Lengauer T, Rarey M (1996) Curr Opin Struct Biol 6:402
14. Hetényi C, van der Spoel D (2002) Protein Sci 11:1729
15. Poornima CS, Dean PM (1995) J Comput Aided Mol Des 9:500
16. Chung SY, Subbiah S (1996) In: Hunter L, Klein TE, Pac Symp Biocomput. World Scientific, Hawaii, USA, pp 126–141
17. Mc Donald IK, Thornton JM (1994) J Mol Biol 238:777
18. Knegtel RMA, Antoon J, Rullmann C, Boelens R, Kaptein R (1994) J Mol Biol 235:318
19. Hubbard SJ, Thornton JM (1993) 'NACCESS': computer program. Department of Biochemistry and Molecular Biology, University College, London
20. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) J Comput Chem 19:1639
21. Tame JR, Dodson EJ, Murshudov G, Higgins CF, Wilkinson AJ (1995) Structure 3:1395
22. Tame JR, Murshudov G, Dodson EJ, Neil TK, Dodson GG, Higgins CF, Wilkinson AJ (1994) Science 264:1578
23. Tame JR, Sleigh SH, Wilkinson AJ, Ladbury JE (1996) Nat Struct Biol 3:998
24. Taylor RD, Jewsbury PJ, Essex JW (2002) J Comput Chem 24:1637
25. Abagyan RA, Totrov M (1997) J Mol Biol 268:678
26. Harel M, Su C-T, Frolow F, Silman I, Sussman JL (1991) Biochemistry 30:5217
27. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ (2003) J Comput Aided Mol Des 17:525
28. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727
29. Oberlin D, Scheraga HA (1998) J Comput Chem 19:71
30. Baysal C, Atilgan AR (2001) Proteins 45:62