MIAX: A System for Assessment of Macromolecular Interaction. 3) A Parallel Hybrid GA for Flexible Protein Docking

Carlos Adriel Del Carpio M. carlos@translell.eco.tut.ac.jp

Atsushi Yoshimori yosimori@translell.eco.tut.ac.jp

Laboratory for Informatics & AI in Molecular and Biological Sciences, Department of Ecological Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441-8580, Japan

Abstract

We propose a parallel hybrid genetic algorithm for flexible protein-protein docking in order to improve the conventional "rigid-body" models to manipulate protein-protein interactions. The proposed hybrid algorithm is a combination of an evolutionary algorithm with a simulated annealing one, yielding a powerful protein-complex conformation-searching engine. Parallelization of the procedure makes possible to reach high algorithm performance, in both, execution times and size of treated monomers and complexes. Knowledge on side chain flexibility is extracted by means of an exhaustive analysis of crystallographic data on proteins and protein complexes. Results demonstrate the competency of the algorithm since comparison of calculated and crystallographic data accounts for a maximum of 2.5Å in RMS difference, including side chain conformation. The system allows routine analysis of this fundamental molecular biology problem important to elucidate bio-macromolecular function in biophysical and biochemical mechanisms involving molecular recognition and interaction, yielding simultaneously clues for designing new proteins and enzymes directed to different purposes.

Keywords: protein-protein interaction, flexible docking, hybrid GA

1 Introduction

Underlying fundamental processes in molecular biology -ranging from gene regulation to enzyme driven reactions- are essentially molecular recognition processes intimately related to macro-molecular interactions.

Classical computer systems directed to model the problem of protein-protein interaction consider it as a "rigid body" interaction problem [2, 3, 5, 9]. This type of models, based generally on geometric complementarity of the monomers, yield complexes of geometrical and structural characteristics close to reported crystallographic conformations, however, they do not succeed in giving better insights into the change of the overall structure of the monomers constituting them, and several times, they fail in giving appropriate account of the residues that are directly involved in the binding process, information which is relevant in modeling this type of interactions. Moreover, the change in conformation on the monomers at interaction are directly related to several biophysical and biochemical processes such as signal production or transmission in which these type of interactions are involved. Therefore, the analysis of these changes that reacting monomers undergo, are of the outmost importance, when analyzing biophysical and biochemical mechanisms of several life sustaining bio-molecular phenomena within organisms, and more important, the acquisition of clues for the design of functional proteins, polypeptides, or other organic molecules in general.

Recently, we have reported on the novel computer system for assessment of macromolecular interaction MIAX [1, 4, 10, 12], giving account of the potential function expressing the interaction of monomers forming a protein complex in solution. In a first approximation to the complexity of this biomolecular process, interaction energies (essentially electrostatic and hydrophobic) were calculated for the system taking into consideration the structural characteristics of the isolated monomers which were kept invariable through the docking procedure (i.e. considering only six degrees of freedom). Here we improve the original system to take into account the contribution of the flexibility of the side chains and the backbone of the protein to the overall docking process and the conformation of the constituting monomers. To this end, the original annealing simulation was embedded within a more general evolutionary algorithm to allow for conformational perturbations in the isolated monomers as they bind to form the new complex molecule. The study comprises analysis of protein crystallographic data to examine tendencies of torsional angles in side chains of isolated proteins and complexes. We found that some amino acid residues possess side chains with torsional angles whose variation is confined to a limited domain while a remarkable flexibility is observed for others. Based on these results side chain torsional angles can be expressed in function of rotamers taking on frequent torsional angles, while increased variability is allowed for highly flexible chains. Furthermore, the program performs automatically perturbations on the backbone of the molecules for accommodating the molecules to form an optimally stable complex molecule.

We present a suite of computational experiments showing the performance of the improved system, which for known crystallographic complexes yields a difference in RMS of 2.5Å as the maximum, considering side chain variation.

These results allow the simulation of the changes on the overall structures of the monomers, going beyond the simple prediction of the complex structure. Because of the parallelization of the algorithm proposed here, which allows almost routine evaluation of protein interaction, we believe that the present work constitutes a first step towards analysis of biophysical, and biochemical mechanisms within organisms in terms of detailed analysis of the interaction of the molecules participating, essentially, accounting for the conformational changes brought about by interaction.

2 Method

Conventional rigid-body protein-protein docking procedures can be described basically by two major steps. The recognition of interaction sites or binding regions on the interacting monomers (by geometrical and/or electrostatic complementarity), followed by refinement of the initial geometry using a suitable potential function expressing the more important terms (hydrophobic, electrostatic, hydrogen bond, etc.) driving the monomers to bind and form the most stable complex [3, 6, 7, 8, 11]. Several studies, however, have shown that conformation of monomers before and after interaction differs in most of the cases [3, 8, 11], which is a natural outcome of the process in which the complex formation is involved (signal processing, production, transmission, etc.).

Here, we have created and implemented a new algorithm to take into account the geometrical changes undergoing monomers at interaction, the system proposed here being an improvement to our original rigid-body docking system, MIAX (Macromolecular Interaction Assessment computer system X).

The algorithm proposed here consists in basically allowing conformation flips of amino acids possessing flexible side chains. Then as consequence of these perturbations, changes in backbone conformation are brought up to account for relaxation of the stress created in the monomers, resulting in the formation of a new stable complex molecule.

To achieve this goal, the algorithm requires the explicit coding of the torsional angles of the side chains of the amino acids constituting the monomers which undergo perturbation.

A method to implement this codification is supplied by the genetic algorithm paradigm (GA), by means of which, the conformation of each side chain in the complex can be recorded as one gene within the chromosomal string representing a determined conformation for the entire complex molecule.

While recording the complete number of sides chains in this way would allow for an exhaustive description of conformation variability, a complete search of the space would be too expensive and almost impossible, even with the most up-to-date high performing computers.

On the other hand, recording the complete number of side chains in the chromosome is not necessary since many of them only have a limited number of torsional angles that must be considered, and furthermore, they take on almost constant values, as found by the analysis described in the next section. Thus, side chains with high flexibility are those driving the most relevant conformational shifts, and are those which are recorded in the chromosomal string.

1) *Analysis of side chain flexibility in proteins*
   We have performed an exhaustive search of all the structures recorded in the latest up-date of the Brookhaven PDB(containing more than 6000 3D strctures) , and made an statistical analysis of the frequency and variability of torsional angles of the side chains for the twenty naturally occurring amino acids. The results of this exhaustive search are summarized in Fig. 1.

   It is evident from Fig. 1 that torsional angles in most side chains take on values that can be assigned to very limited variation domains.

   Similarly evident is the fact that torsional angles for amino acids such as GLN(χ^3), TRP(χ^2), ARG(χ^4) and ASN(χ^2) are those presenting the highest variability.

2) *Encoding side chains as rotamers*
   The analysis in the previous section allows a simple codification scheme to represent a determined conformation for the complex under formation, as a string of torsional angles of side chains and backbones, as schematized in Fig.2. The relative position among monomers constituting the complex is also expressed within the chromosomal structure. This information comprises the inter-monomer distance and the Euler angles specifying their spatial position respective to one another.

3) *The GA-SA Hybrid Algorithm*
   While codification of the entire solution of the problem (the chromosome representing the complex conformation) is achieved as described in the previous section, the optimal solution (i.e., conformation of the most stable complex) is obtained by searching the space using the three fundamental operations of the GA (selection, crossover and mutation). Because of the exhaustiveness required in the search of the hyper-space of solution, avoiding traps in relative minima is performed by the original simulated annealing algorithm[1, 12]. Accordingly, the searching engine consists in evolving chromosomes that are further minimized when the objective function (or penalty function, in this case the energy of the specific complex conformation) is computed, generating in this way only the best conformers constituting the population of chromosomes at each generation of the GA.

   Energy of a particular complex conformation is computed as described elsewhere[1, 12], and can be summarized in the following expression:

$$\Delta G = E_{hy} + E_{hb} + E_{elec} + E_{tor} + E_{desol} \tag{1}$$

   where E_{hy}, stands for the hydrophobic interaction energy, (defined here as the cavity formation energy for the complex and the van der Waals interaction energy), E_{hb}, the hydrogen bond energy, E_{elec} is the electrostatic energy, and E_{desol} is the desolvation energy for the molecule. Finally E_{tor} is the term expressing the energy due to intra-molecular conformational changes. The calculation of each term for the complex for a given conformation is described in detail in [1].

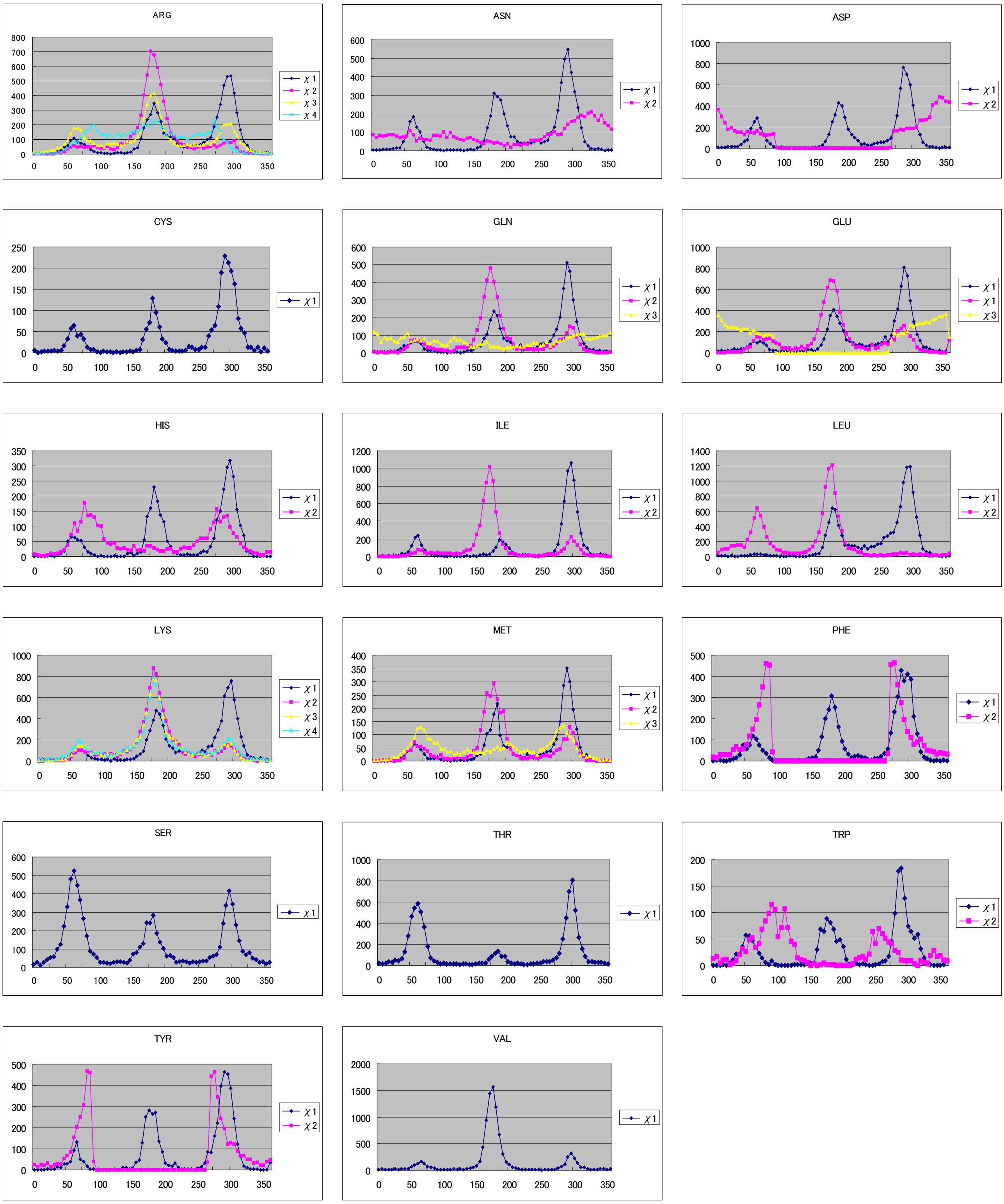

Figure 1: Side chain torsional angles variability in seventeen naturally occurring amino acid residues. (χ^n: n th torsion angle in the side chain)

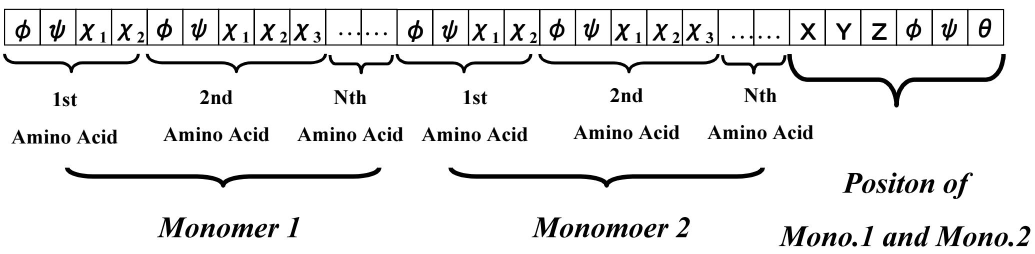

Figure 2: Codification of a protein complex comformer as a chromosome in the hybrid GA. (ϕ, ψ: backbone torsional angles, x_n: n th torsional angle of a residue, $X, Y, Z, \phi, \psi, \theta$: coodinates to control relative positions of the monomers)

4) *Simulation and the Parallel Algorithm*
Higher performance efficiency of the algorithm so far described is achieved by parallelization of the searching engine. The parallelization scheme considered here consists in dividing populations of chromosomes representing complex conformations according to the number of processors available in the parallel computer system. Recombination of chromosomes from different processors after a certain number of generations adds robustness and diversity to the search engine.

3 Results and Discussion

Representation of side chain conformations as rotamers with specific variation intervals (i.e. those with single or double torsional angles, as shown in Fig. 1) allows a simple operation of flipping the angle values, and adopting those which cause the least steric strain of the side chain within its environment. Shifts from the intervals of variation are kept to a minimum and are adopted only when steric strains prevail unless a variation of the torsional angle value. Torsional angles of highly flexible side chains are those dictated by the simulated annealing minimization and can cover a wide range of values.

The effectiveness of the improvement to the rigid-body docking of MIAX is exemplified with a suite of computer experiments of flexible docking of polypeptides.

Data for the experiments were taken from experimentally known complexes, or monomers known to interact and form complexes whose structures are recorded in the Brookhaven PDB. For cases when only the crystal complexes are known, the monomers are separated, and perturbation of the side chains is performed in order to have the search start from a conformation different to that at the interaction position.

The first step in the new flexible docking module in MIAX is similar to that of the "rigid-body" case, which consists in searching the initial position by means of a geometric complementarity finding algorithm like FTDOCK [3, 9].

Table 1, summarizes the complexes utilized in the work, each referenced by their PDB code and showing the initial RMSd (after separation of the complexes, perturbation of the side chains, and placing the monomers at positions dictated by complementarity analysis by FTDOCK [3, 9]). And the last column shows the minimum RMSd obtained by the flexible docking module in MIAX.

1) For 1cka (oncogene/peptide complex), the initial position and the final position compared with the crystal structure (Fig. 3), as well as the energy and RMSd variation through the evolutionary process are shown in Fig. 4. A rigid-body docking (results not shown here) yields a complex at no less than the initial RMSd, for the backbone of the protein alone. The flexible docking however is able to bring down this difference to a much lower value, but also considering atoms constituting the side chains.

2) For 9hvp (HIV_1-Protease complex), the initial RMSd decreases from 2.4 A to 0.3 A by flexible

Table 1: Result from Docking Simulation.

Complex (PDB Code)	Chain	Amino Acid		Initial RMS[Å]	Minimum RMS[Å]
1cka	a, b	57	10	3.4	2.4
9hvp	a, b	99	99	2.4	0.3
2mhb	a, b	141	146	1.3	0.8
2utg	a, b	70	70	1.1	0.6
2pab	a, b	127	127	5.8	2.5

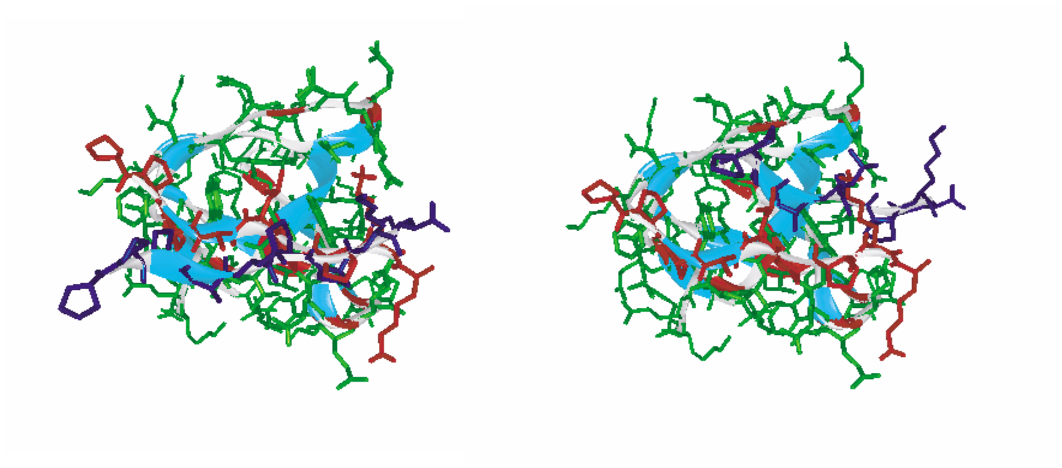

Figure 3: MIAX output for complex 1cka. RIGHT: Starting position found by FTDOCK (black), relative to the crystal structure (brown). LEFT: Minimized structure output by MIAX (black), relative to the crystal structure (brown). (RMSd: 2.4Å including side chain atoms)

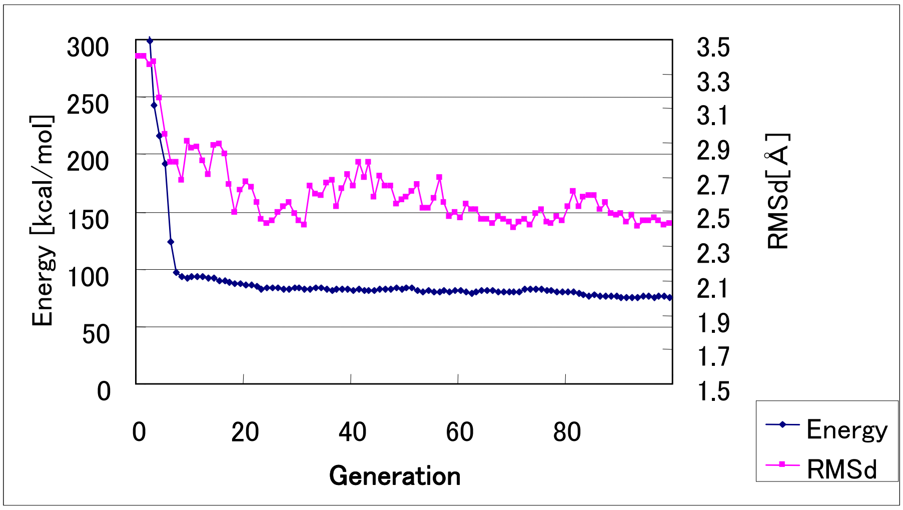

Figure 4: Minimization process by the GA-SA hybrid algorithm for complex 1cka

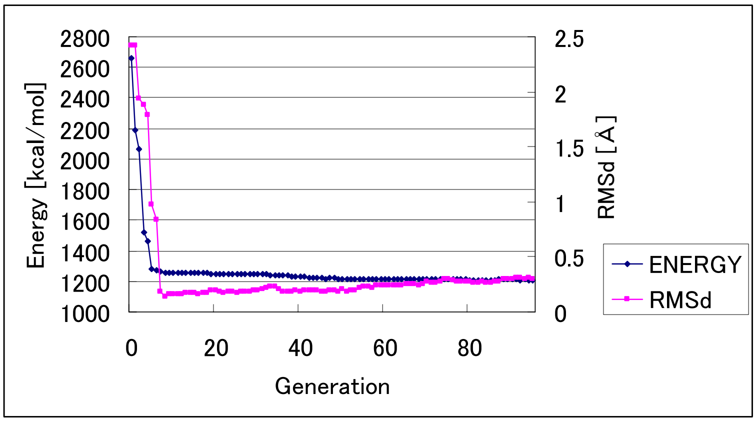

Figure 5: Minimization process by the GA-SA hybrid algorithm for complex 9hvp

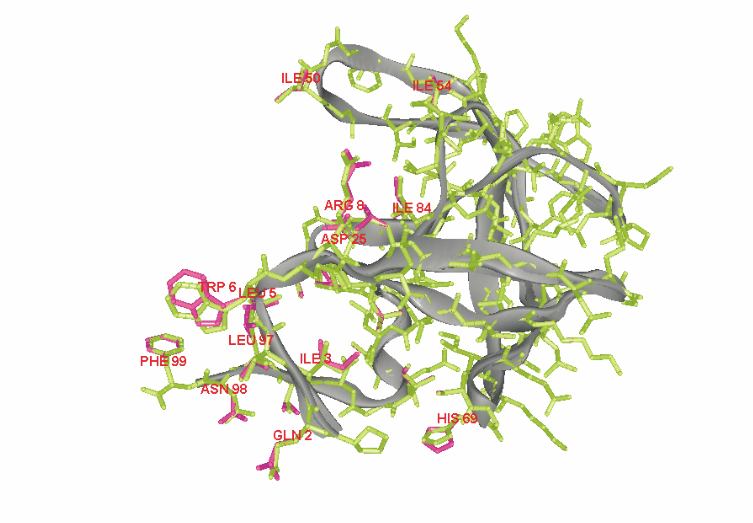

Figure 6: Comparison of the interaction interface the homo-dimer, 9hvp (Green: Crystral structure, Purple: MIAX output)

docking. Variation of the RMSd and the energy as the minimization process proceeds as depicted in Fig.5. The interface at interaction of this homodimer is illustrated in Fig.6, where those amino acids with high flexibility and directly involved in the interaction can be identified. From this plot it is evident that the amino acids labeled TRP6 PHE99 and ARG8 have difficulty in adopting the conformation in the crystal structure, due to the wide range of angles that can be adopted by every torsional angle in the side chains. Differences are, nevertheless, negligible, if the resolution of the crystal structure is taken into consideration.

Similar plots for 2mhb and 2utg are shown in Fig. 7 (Rms=0.8Å) and Fig. 8 (Rms=0.6Å) respectively. It must be remarked the high correlation between the enrgy and the RMSd of complexes through the artificial evolution process in MIAX.

4 Conclusions

Here, we address the problem of flexible protein-protein docking, as an improvement to our previously reported general system for macromolecular interaction assessment MIAX. The present study has

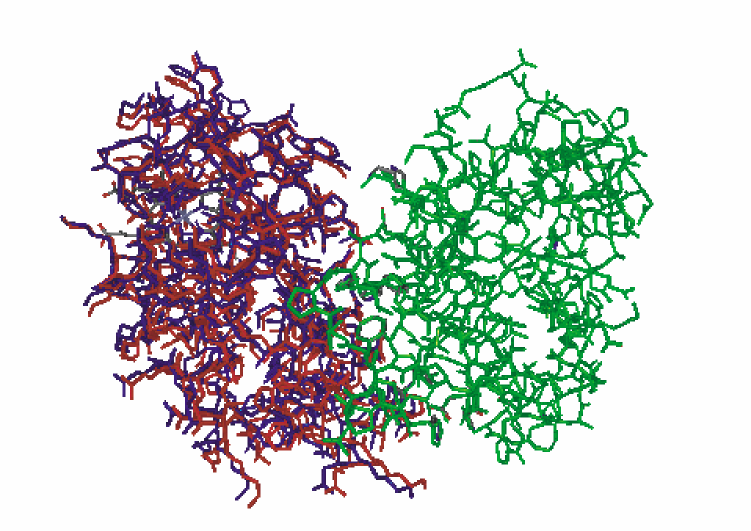

Figure 7: Superposition of the automatic docking (black) and the crystal structure (brown) for complex 2mhb (see Table 1).

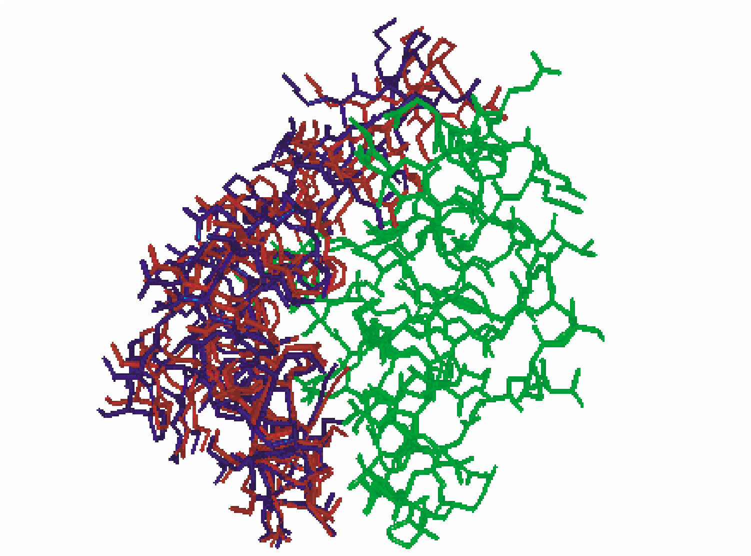

Figure 8: Superposition of the automatic docking (black) and the crystal structure (brown) for complex 2utg (see Table 1).

two important objectives to accomplish, the first is related essentially to the demonstration of the capability of the solvation effect term in our original system [1], to drive the monomers into the optimal interaction site, both in terms of relative positions and also in terms of conformational changes that they undergo as they bind.

This is closely related to the distinction that the system must make among several alternatives of binding that are output by complementarity analysis of the interacting monomers, since programs such as FTDOCK[3,9], and also others developed in our laboratories[12], output a series of binding region candidates. This first goal has been achieved, since the RMSd of the molecule decreases proportionally to the decrease in energy of the entire molecular system, as shown in all the examples presented in the previous section. This is further corroborated by the almost insignificant difference in RMSd achieved by the improved system, which considers not only the relative position of the molecules respect each other, but the changes in conformation of the backbone and side chains.

The system is able to identify those amino acids with high flexibility in the monomers, and their contributions to binding for the unbound and bound states. The refinement of the monomers conformation at the interaction interface determine the most relevant residues interaction at binding , and this signifies a step forward in modelling protein interactions given the inclusion of conformational flexibility and the treatment of solvation so far proposed in our previous work.

A second goal, which has been only partially achieved, and which still requires further analysis, is that related to the overall effect on the molecules as the interaction of proteins proceeds. Although the problem is rather complex, and because our methodology is more oriented to obtain the optimal complex configuration for a determined complex system given the conformation of the monomers. Some hypothesis can be made about fundamental changes brought about by protein-protein interaction, that can be related to functions corresponding to biophysical or biochemical mechanisms in which they are involved.

The improvement in the form of a new module developed in the present study sets the general framework for this type of evaluations, since the codification of the solution adopted by the GA takes into consideration all the possible changes in the molecule, and not only those at the interface or binding region. Therefore, comparison of the conformations adopted by the conformers within the complex with those of the the original monomers is straightforward, making possible the direct correlation between structural change and the function that may be accomplished by the interacting molecules when they form the new molecule. Changes in monomer backbones may be the ones related to interchange among allotropic states of the monomers, because of the significant changes in conformation that occur along with them through the interaction process.

Acknowledgement

We are very grateful to Mr. Masaki Kobayashi, president of KLIMERS Corp. for his enormous efforts in supporting the present research both economically, supplying the parallel system on which MIAX runs, and experimentally testing the results obtained theoretically.

References

[1] Del Carpio, C.A. and Yoshimori A., MIAX: A novel system for assessment of macromolecular interaction in condensed phases. 1) description of the interaction model and simulation algorithm, *Genome Informatics*, 10:3–12, 1999.

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

[3] Gabb, H.A., Jackson, R.M., and Sternberg, M.J.E., Modelling protein docking using shape complementarity, electrostatics and biochemical information, *J. Mol. Biol.*, 272:106–120, 1997.

[4] Hara, T. and Del Carpio, C.A., A simulated annealing algorithm for geometrical assessment of macromolecular hydrophobic interaction, *Genome Informatics*, 9:380–381, 1998.

[5] Hou, T., Wang, J., Chen, L., and Xu, X., Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu serch., *Protein Engineering*, 12:639–647, 1999.

[6] Ippolito, J.A., Alexander, R.S., and Christianson, D.W., Hydrogen bond stereochemistry in protein structure and function, *J. Mol. Biol.*, 215:457–471, 1990.

[7] Jackson, R.M., Gabb H.A., and Sternberg, M.J.E., Rapid refinement of protein interfaces incorporating solvation: Application to the Docking Problem, *J. Mol. Biol.*, 276:265–285, 1998.

[8] Jiang, F. and Kim, S., "Soft Docking": matching of molecular surface cubes, *J. Mol. Biol.*, 219:79–102, 1991.

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

[10] Matsumoto, H., Yoshimori, A., Kojima, T., and Del Carpio, C.A., Analysis of sense-antisense peptide interaction using molecular mechanics, *Peptide Chemistry*, 173–176, 1996.

[11] Meyer, M., Wilson, P., and Schomburg, D., Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking, *J. Mol. Biol.*, 264:199–210, 1996.

[12] Yoshimori, A. and Del Carpio, C.A., A novel system for assessment of macromolecular interaction in condensed phases. 2) interaction site inference by molecular shape and electrostatic complementarity, *Genome Informatics*, 10:263–264, 1999.