Biomolecular Structure Prediction Stochastic Optimization Methods**

By *Alexander Schug*,* *Bernhard Fischer*, *Abhinav Verma*, *Holger Merlitz*, *Wolfgang Wenzel* and *Gerd Schoen*

Biomolecular structure prediction remains an important challenge to biophysical chemistry. We recently developed an all-atom free energy forcefield (PFF01) for protein structure prediction with stochastic optimization methods. We review recent studies, which demonstrated all-atom folding of several proteins and summarize recent progress for in-silico high-throughput screening strategies for rational drug design, which are also based on the use of stochastic optimization methods to determine the conformation of the receptor-ligand complex.

1. Introduction

The understanding and prediction of biomolecular structure plays an important role in many diverse applications ranging from biology to nanobiotechnology. *Ab-initio* protein tertiary structure prediction (PSP) and the elucidation of the mechanism of the folding process are among the most important outstanding problems of biophysical chemistry.[1] Is a three-dimensional structure avaiable, *in-silico* high-throughput screening is rapidly becoming a viable tool for rational drug design. Here we briefly review recent progress made in both of these areas.

2. Protein Structure Prediction

It has been one of the central paradigms of protein folding that proteins in their native conformation are in thermodynamic equilibrium with their environment.[2] Exploiting this characteristic the structure of the protein can be predicted by locating the global minimum of its free energy surface without recourse to the folding dynamics, a process which is potentially much more efficient than the direct simulation of the folding process.

We have recently demonstrated a feasible strategy for all-atom protein structure prediction[3,4,5] in a minimal thermodynamic approach. We developed an all-atom free-energy forcefield for proteins (PFF01),[5] which permitted the reproducible and predictive folding of four proteins, the 20 amino acid trp-cage protein (1L2Y),[3,6] the structurally conserved headpiece of the 40 amino acid HIV accessory protein (1F4I)[4,7] and the sixty amino acid bacterial ribosomal protein L20.[8]

2.1. Methods

The all-atom free-energy protein forcefield (PFF01) models the low-energy conformations of proteins with minimal computational demand.[4,5] In the folding process at physiological conditions the degrees of freedom of a peptide are confined to rotations about single bonds. The Lennard Jones parameters for potential depths and equilibrium distance) depend on the type of the atom pair and were adjusted to satisfy constraints derived from as a set of 138 proteins of the PDB database. The non-trivial electrostatic interactions in proteins

[*] *Dr. A. Schug, Prof. B. Fischer, Dr. H. Merlitz, Dr. W. Wenzel*
Forschungszentrum Karlsruhe
Institute for Nanotechnology
P.O. Box 3640, D-76021 Karlsruhe, Germany
Prof. A. Verma
Forschungszentrum Karlsruhe
Institute for Scientific Computing
P.O. Box 3640, D-76021 Karlsruhe, Germany
Prof. Dr. G. Schoen
Institut für Theoretische Festkörperphysik
Universität Karlsruhe
D-76021 Karlsruhe, Germany

[**] *We thank the BMBF, the Deutsche Forschungsgemeinschaft (grant WE 1863/11-1, WE 1863/10-2) and the Kurt Eberhard Bode Stiftung for financial support. Part of these calculations was performed on the KIST Materials Simulation Laboratory Supercomputer Cluster.*

are represented via group-specific dielectric constants and interactions with the solvent in solvent accessible surface model. Hydrogen bonds are described via dipole-dipole interactions included in the electrostatic terms and an additional short range term for backbone-backbone hydrogen bonding.[5]

The low-energy free energy landscape of proteins is extremely rugged due to the comparatively close packing of the atoms in the native structure. Suitable optimization methods must therefore be able speed the simulation by avoiding high energy transition states, adapt large scale move or accept unphysical intermediates. Here we report on four different optimization methods, the stochastic tunneling method,[9] the basin hopping technique,[10,11] the parallel tempering method [12] and a recently employed evolutionary technique.[8]

2.2. Results

Using the PFF01 forcefield we simulated 20 independent replicas of the 20 amino acid trp-cage protein[13,14] (pdb code 1L2Y) with a modified versions of the stochastic tunneling method.[9,3] Six of 25 simulations reached an energy within 1 kcal/mol of the best energy, all of which correctly predicted the native experimental structure of the protein (see Fig. 1 (left)). We find a strong correlation between energy and RMSD deviation to the native structure for all simulations. The conformation with the lowest energy had a backbone root mean square deviation of 2.83 Å.

We also folded this protein with the parallel tempering method,[6] where the best final structure had a RMSB deviation of just 2.01 Å and with the basin hopping technique,[3] where the lowest six of 20 simulations converged to the native structure.

We then applied the modified basin hopping methods to fold the structurally conserved 40-amino acid headpiece of the HIV accessory protein.[4] We performed twenty independent simulations and found the lowest five to converge to the native structure.[8] The first non-native decoy appears in position six, with an energy deviation of 5 kcal/mol and a significant RMSB deviation. The good agreement between the folded and the experimental structure is also evident from Figure 1 (center), which shows the secondary structure alignment of the native and the folded conformations. We also folded the HIV accessory protein with adapted parallel tempering method[7] using 20 replicas to 2.46 Å backbone root mean square (RMSB) deviation to the experimental structure. Considering the ensemble of final conformations, we find many structures closely resembling the native conformation.

For the 60 amino acid bacterial ribosomal protein L20 (pdb-code 1GYZ) we experimented with the evolutionary technique described in the methods section. Starting from a seed population of random structures we performed the folding simulation in three phases: (1) generation of starting structures of the population, (2) evolutionary improvement of the population and (3) refinement of the best resulting structures to ensure convergence. Again the best conformation had approached the native conformation to about 4.6 Å RMSB deviation. In total six of the lowest ten conformations approach the native structure, while four others misfolded.

2.3. Summary

Since the native structure dominates the low-energy conformations arising in all of these simulation, our results demonstrate the feasibility of all-atom protein tertiary structure prediction for three different proteins ranging from 20 60 amino acids in length with a variety of different optimization methods. The free energy approach thus emerges as viable trade-off between predictivity and computational feasibility. While sacrificing the folding dynamics, a reliable prediction of its terminus, the native conformation -- which is central to most biological questions -- can be achieved.

3. Receptor Ligand Docking

Virtual screening of chemical databases to targets of known three-dimensional structure is developing into an increasingly reliable method for finding new lead candidates in drug development. Both better scoring functions and novel docking strategies contribute to this trend, although no completely satisfying approach has been established yet. This is not surprising since the approximations which are needed to achieve a reasonable screening rates impose significant restrictions on the virtual representation of the physical system. Relaxation of these restrictions, such as permitting ligand or receptor flexibility, potentially increase the reliability of the scoring process, but come at a high computational cost.

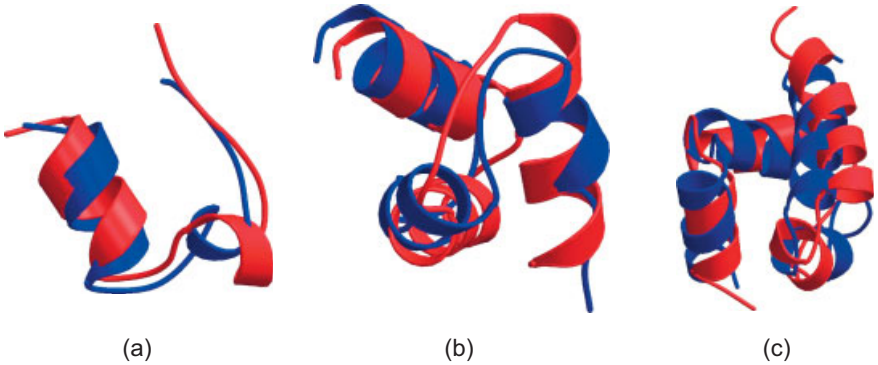

Fig. 1. Overlay of the native(red) and folded (blue) structures of trp-cage protein, the HIV accessory protein and the bacterial ribosomal protein L20.

The limitations of presently available computational resources and the large number of possible ligands enforce severe approximations in the representation of receptor and ligand, and their interactions. Significant computational efficiency is gained, when the protein receptor is assumed to be rigid in the docking process; for this reason many tests of screening functions[15] and virtually all large scale computational screens presently rely on a rigid-receptor conformation. On the other hand, direct comparison between ligand-free and complexed crystal structures often demonstrate a significant ligand-induced alteration of the receptor structure.

While ligand flexibility is now routinely considered in many atomistic in-silico screening methods, accounting for receptor flexibility still poses significant challenges.[16,17,18] The thymidine kinase receptor is a useful benchmark system for the evaluation of screening methods, because not just one, but several substrates are known and characterized in the their binding mode. Here we use this system as a prototypical example to document the shortcoming of rigid receptor screens, independent of the particular choice of the receptor conformations. We then present screens using *FlexScreen*,[19,20] a recently developed all-atom screening tool based on the stochastic tunneling method to screen a subset of up to 10000 ligands of the NCI-Open database considering receptor side-chain flexibility.

In the following we first describe the two main ingredients to *all-atom in-silico screening*: the docking tool *FlexScreen* and the parameterization of the scoring function that approximates the binding energy of the ligand to the receptor. Next we present the results of several screens of 10000 ligands of the NCI database against specific rigid receptor conformations and introduce a scoring scheme that quantifies the quality of a particular screen. Finally we perform a screen with a *flexible receptor* and discuss its advantages compared to the previous rigid receptor screens.

3.1. Methods

There are two major ingredients to an all-atom in-silico screening method: (1) a scoring function that approximates the binding energy (ideally the affinity) of the receptor-ligand complex as a function of the conformation of this complex. and (2) an efficient optimization method that is able to locate the binding mode of a given ligand to the receptor as the global optimum of the scoring function. In a database screen, all ligands are thus assigned an optimal score which is then used to sort the database to select suitable ligands for further investigations.

The screens in this investigation were performed with *FlexScreen*, an all-atom docking approach[19,20] based on the stochastic tunneling method.[9] This method was shown to be superior to other competing stochastic optimization methods[19] and had performed adequately in a screening of 10000 ligands to the active site of dihydrofolate reductase (pdb code 4 dfr[21]), where the known inhibitor (methotrexate) emerged as the top scoring ligand.[20]

Many different scoring functions have been proposed in recent years[22] and no clear consensus has emerged to date on the superiority of physics-based or knowledge based approaches. In this investigation we employed a simple, first-principle scoring function, which proved successful in a prior investigation of the dihydrofolate reductase receptor.[20]

3.2. Results

We investigated the degree of database enrichment of 10000 compounds, randomly chosen from the nciopen3D database,[25] and 10 known substrates when docked to the X-ray TK receptor structure, which was experimentally determined in complex with one of the substrates, dt (deoxythymidine, pdb entry 1ki2[26]). In this screen 5353 ligands attained a stable conformation with negative affinity within the receptor pocket. Figure 2 shows the number of ligands as a function of affinity and highlights the rank of the known TK substrates in the screen. Three structurally similar substrates, including the ligand associated with the receptor conformation, are ranked with very high affinity. This result demonstrates that docking method and scoring function are adequate to approximate the affinity of these ligands to the receptor. Four further ligands (idu, acv,gvc, pcv, for a detailed description of TK and its substrates we refer to[15]) docked badly, three further ligands did not dock at all according to the criteria above. Repeating the docking simulations for these ligands did not substantially improve their rank in the database, eliminating inaccuracies of the docking algorithm as the source for this difficulty.

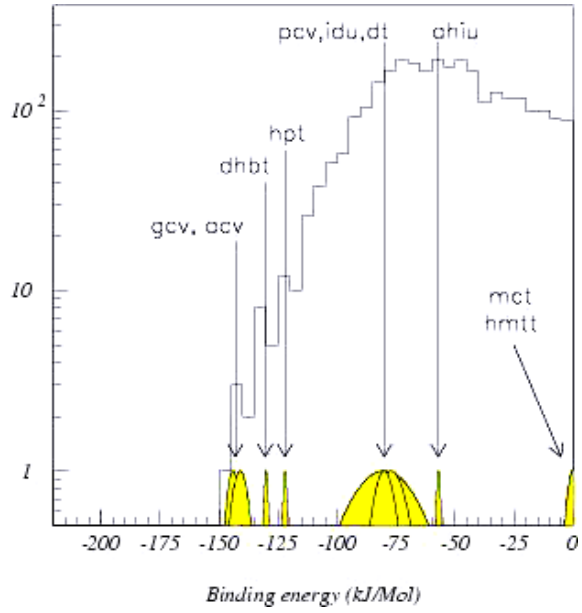

Fig. 2. Histogram of the affinities of the 6284 docked ligands (see text) to the rigid receptor conformation complexed with ganciclovir. The count is plotted on a logarithmic scale. While some known substrates, in particular the substrate corresponding to the receptor conformation, score well compared to the ligands of the randomly selected database, several other fail to achieve good scores.

Table 1. Ranking of the TK substrates in a screen of 10000 randomly chosen ligands of the nciopen3D database. The top row designates the crystal structure of the receptor that was used in the screen, the last column indicates the results of a flexible receptor screen. (nd = not docked)

Substrate	1kim	1ki2	1ki3	1e2h	flex
acv	719	9	22	2048	199
ahiu	nd[c]	nd	nd	nd	2673
dhbt	4	104	118	38	13
dt	5	1310	2576	2779	681
gcv	3351	78	15	4516	57
hmtt	nd	nd	nd	nd	656
hpt	6	152	266	36	148
idu	515	2436	3272	2913	1365
mct	nd	6074	nd	nd	247
pcv	4845	952	4	4739	1656
Score:	3751	3705	4575	1926	4999

The resulting ranks of this screen are summarized in Table 1 (second column), which displays the rankings of the 10 substrates. Three were ranked within the first 1 %, 6 were ranked among the first 10% of the database, respectively. This enrichment rate is comparable to the results of other scoring functions that were previously investigated for this system, but the overall performance is disappointing.[15]

Inspection of the crystal structures of the different receptor-ligand complexes reveal differences in the conformation of some side groups inside the receptor pocket, depending on the docked substrate. This is a well known fact, but it is often assumed that the impact of these conformational variations on the ranking accuracy is moderate. We therefore repeated the screening with the X-ray structure of TK in complex with the substrate gcv (ganciclovir, pdb entry: 1ki2[26]), which had scored particularly bad in the original screen. The results are shown in Table 1 (third column). Now, gcv was ranked within the leading 1 % of the database, but dt, formerly ranked on position 5, dropped to 1310. The same procedure was then repeatedw ith TK in complex with pcv (penciclovir, pdb entry: 1ki3), which raised its rank from 4845 to 4 (fourth column of Tab. 1).

For comparison purposes, we also performed a screen of the ligand free X-ray structure of TK (pdb entry: 1e2h[27]), which would most likely be used in a screen if no substrate was known. In this screen the receptor is unbiased to any of the substrates, which results in a dramatic loss of screening performance. As shown in column 7 only two ligands scored reasonably well (within the upper 10 % of the database), all others would be discarded by any rational criterion as possible lead candidates.

Next we performed a flexible receptor screen against the same database. We identified the critical amino acid side chains and introduced 23 receptor degrees of freedom into the structure 1ki2, i.e. dihedral rotations of the amino acids His13(2), Gln76(3), Arg173(4), Glu176(4), Tyr52(3), Tyr123(3) and Glu34(4). The numbers in brackets indicate the degrees of freedom for each sidechain. Each step in the stochastic search now consisted of an additional random rotation for one receptor degree of freedom. The results of this screen are summarized in Figure 3, the scores of the individual substrates listed in the column labeled 'flex' in the table. The figure demonstrates that in contrast to all rigid receptor screens now all substrates dock to the receptor. As expected, the number of false positives also increases, because a flexible conformation of the receptor reduces the bias of the screen against the known substrates. It must be noted that the accuracy of the flexible receptor screen is lower than that of the rigid receptor screens (with the same number of function evaluations) because the number of degrees of freedom has increased. The increased fluctuations in the flexible screen can be best seen for acv, where the optimization method failed to locate the global optimum of the affinity (as independently obtained in a longer scree for just this ligand). We are presently developing algorithms to only selectively move the sidechains to reduce the computational effort in the flexible receptor screen.

3.3. Summary

To quantitatively compare different screens against the same ligand database, which used different receptor geometries, scoring functions or docking methods, it is sensible to assign an overall score to each screen which rates its performance.[28] We computed such a "score" for the entire screen from the ranks of the docked known substrates among the N=1000 best ligands. This score is computed as the sum of N-P where P is the rank of the known substrate and shown in

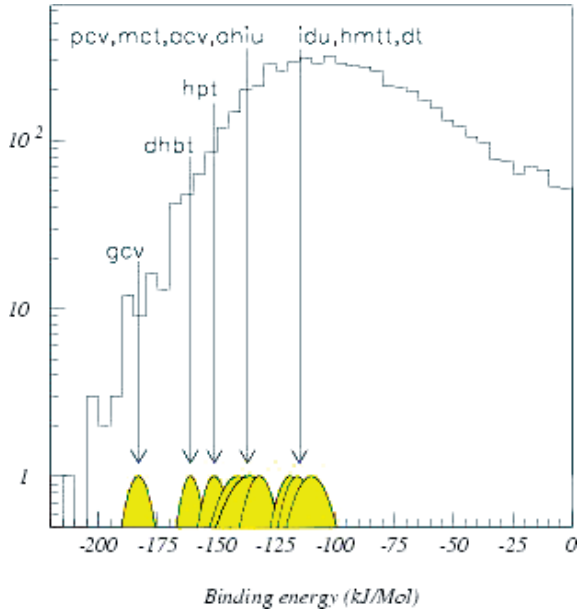

Fig. 3. Histogram of the affinities of the docked ligands in the flexible receptor screen.

the bottom row of Table 1. A substrate ranking in the top of the screen contributes a score of 1000 to the sum, a badly ranked substrate comparatively little. Because the best *N* ligands are evaluated, screens which dock many known substrates with moderate rank may have comparable scores with screens which perform perfectly for one substrate, but fail for all others. For the rigid receptor screens performed here the scores for the entire screen ranged from between 1926 for the screen against 1e2h, the ligand free X-ray structure of TK, to 4575 (1ki3, X-ray structure of TK in complex with pcv), which was arguably the best performing screen of all receptor conformations. Despite the increase in the number of false positives the overall score of the flexible receptor screen (4999) was better than that of any rigid receptor screens.

Our results offer a good demonstration that the ranking of known substrates can strongly depend on the particular receptor structure used for the screen. Binding energy and rank of a given substrate differ significantly depending on the receptor conformations. Our data demonstrate that this variability in rank is not, in general, a shortcoming of either scoring function nor docking methodology. Using a fixed three-dimensional structure of the receptor that is suitable for a single ligand introduces a significant bias in the overall scoring of the entire database. As a consequence, differences in the enrichment ratio for different scoring functions[15] may depend more on the suitability of the receptor conformation and environment than on the quality of the scoring function.

These findings suggest the importance of the consideration of a flexible binding pocket to obtain a better unbiased scoring of high-affinity ligands. The results of the flexible receptor screen reported here suggest that better accuracy of the scoring process can be achieved when receptor flexibility is considered. Ultimately only the routine use of accurate scoring techniques for flexible receptors, such as *FlexScreen*, will ameliorate this problem. The results presented here demonstrate that such screens will become feasible with present day computational resources in the near future.

[1] D. Baker, A. Sali, *Science* **2001**, *294*, 93.
[2] J. Schonbrunn, W. J. Wedemeyer, D. Baker, *Curr. Op. Struc. Biol* **2002**, *12*, 348.
[3] N. Go, H. A. Scheraga, *Macromol.* **1976**, *9*, 535.
[4] P. Ulrich, W. Scott, W. W. F. van Gunsteren, A. E. Torda, *Proteins, SF&G* **1997**, *27*, 367.
[5] C. D. Snow, H. Nguyen, V. S. Pande, M. Gruebele, *Nature* **2002**, *420*, 102.
[6] C. Simmerling, B. Strockbine, A. Roitberg, *J. Am. Chem. Soc.* **2002**, *124*, 11258.
[7] C. B. Anfinsen, *Science* **1973**, *181*, 223.
[8] Z. Li, H. Scheraga, *Proc. Nat. Acad. Sci. U.S.A.* **1987**, *84*, 6611.
[9] A. Schug, T. Herges, W. Wenzel, *Phys. Rev. Letters* **2003**, *91*, 158102.
[10] T. Herges, W. Wenzel, *Phys. Rev. Letters* **2004**, *94*, 018101.
[11] T. Herges, W. Wenzel, *Biophys. J.* **2004**, *87*, 3100.
[12] A. Schug, T. Herges, W. Wenzel, *Europhyics Lett.* **2004**, *67*, 307.
[13] A. Schug, T. Herges, W. Wenzel, *Proteins* **2004**, *57*, 792.
[14] A. Schug, T. Herges, W. Wenzel, *J. Am. Chem. Soc.* **2004**, *126*, 16736.
[15] H. Gouda, et.al. Bi*ochemistry* **1992**, *40*, 9665.
[16] U. Mayor, et. al., *Nature* **2003**, *421*, 863.
[17] T. Herges, H. Merlitz, W. Wenzel, *J. Ass. Lab. Autom.* **2002**, *7*, 98.
[18] R. Abagyan, M. Totrov, *J. Molec. Biol.* **1994**, *235*, 983.
[19] T. Herges, A. Schug, B. Burghardt, W. Wenzel, *Intl. J. Quant. Chem.* **2004**, *99*, 854.
[20] F. Avbelj, J. Moult, *Biochemistry* **1995**, *34*, 755.
[21] D. Eisenberg, A. D. McLachlan, *Nature* **1986**, *319*, 199.
[22] K. A. Sharp, A. Nicholls, R. Friedman, B. Honig, *Biochemistry* **1991**, *30*, 9686.
[23] W. Wenzel, K. Hamacher, *Phys. Rev. Lett.* **1999**, *82*, 3003.
[24] A. Nayeem, J. Vila, H. Scheraga, *J. Comp. Chem.* **1991**, *12*, 594.
[25] J. P. Doye, D. Wales, *J. Chem. Phys.* **1996**, *105*, 8428.
[26] G. J. Geyer, *Stat. Sci.* **1992**, *7*, 437.
[27] K. Hukushima, K. Nemoto, *J. Phys. Soc. Japan* **1996**, *65*, 1604.
[28] H. Merlitz, W. Wenzel, *Chem. Phys. Lett.* **2002**, *362*, 271.
[29] H. Merlitz, B. Burghardt, W. Wenzel, *Chem. Phys. Lett.* **2003**, *370*, 68.
[30] U. Hansmann, Y. Okamoto, *J. Comput. Chem* **1997**, *18*, 920.
[31] U. Hansmann, *Eur. Phys. J. B* **1999**, *12*, 607.
[32] C. Lin, C. Hu, U. Hansmann, *Proteins* **2003**, *53*, 436.
[33] S. Kirkpatrick, C. Gelatt, M. Vecchi, *Science* **1983**, *220*, 671.
[34] J. Schneider, I. Morgenstern, J. Singer, *Phys. Rev.* E **1998**, *58*, 5085.
[35] A. Schug, A. Verma, T. Herges, K. H. Lee, W. Wenzel,, *ChemPhysChem* **2005** (in press).
[36] J. W. Neidigh, R. M. Fesinmeyer, N. H. Anderson, *Nature Struct. Biol.* **2002**, *9*, 425.