Stochastic optimization methods for structure prediction of biomolecular nanoscale systems

T Herges, A Schug, H Merlitz and W Wenzel

Forschungszentrum Karlsruhe, Institut für Nanotechnologie, Postfach 3640, D-76021 Karlsruhe, Germany

E-mail: Wolfgang.Wenzel@int.fzk.de

Online at stacks.iop.org/Nano/14/1161

Abstract
The development of simulation techniques that can elucidate the function of biomolecular nanodevices is still in its infancy. In this paper we summarize our approach to the investigation of structural properties of biomolecular systems with stochastic optimization methods. We briefly review the stochastic tunnelling method and summarize applications in two important areas of biomolecular structure prediction: protein folding and protein–ligand association.

1. Introduction

Biomolecular structure prediction remains one of the main outstanding problems of theoretical biophysical chemistry [1]. In particular, with application to problems in nanotechnology [2, 3], where biomaterials interface with inorganic substrates [4], few theoretical methods to elucidate structure and dynamics with atomic resolution are currently available [5]. The direct simulation of association processes or functioning nanoscale devices is complicated by the fact that interesting processes in these systems often occur at timescales that are very long compared with a typical time step of the simulation [6].

For systems in thermodynamic equilibrium with their environment, however, there are often alternatives to the direct simulation of the process, provided a free-energy functional of the system is known. If this is the case, structural questions can be addressed orders of magnitude faster by sacrificing the information about the process by which the thermodynamically stable conformation is reached [7]. Using stochastic optimization methods [8–10], one can determine the global optimum of the free-energy surface (FES) of the system without recourse to the folding dynamics.

Even dynamic questions can be addressed if the low-energy local minima of the FES can be resolved [11]. An (adiabatic) functioning nanomachine typically switches between such low-energy minima [3], which change their respective energies as a function of an external field that triggers the switch of the machine.

This approach requires the development of (i) accurate force fields for the system in question and (ii) efficient optimization methods to resolve the low-energy minima of the FES. Here we review recent progress we have made regarding both points in the context of two important applications of biomolecular structure prediction.

First we address the prediction of the three-dimensional, tertiary structure of proteins on the basis of their amino acid sequence (protein structure prediction—PSP). Experimental methods to determine the sequence of a particular protein have made enormous progress, resulting in a pool of several hundred thousand sequenced proteins for various organisms. Experimental methods for protein structure determination are orders of magnitude more involved and more expensive than sequencing techniques. Although their number is steadily growing, the protein database (PDB), currently contains about 20 000 spatially resolved structures [12]. Theoretical methods for PSP may be helpful to close this gap, but accurate theoretical methods that would permit a routine prediction of this structure remain elusive, in particular at the *ab initio* level [13]. In particular for application in the area of nanobiotechnology, where peptides and proteins often do not occur in their natural but in a technologically altered environment, theoretical methods to describe their structure may be valuable.

In the following we review our strategy for developing an all-atom protein force field with an implicit solvent model to describe such molecules and stochastic optimization methods that are particularly well suited to determine their low-energy configuration. While the former is directly relevant for applications dealing with proteins or peptides, complex optimization problems play a role in many areas of nanotechnology and the methods reviewed below may be of use for such problems.

Secondly we report on the application of stochastic optimization methods to biomolecular switching with pharmaceutical applications. The goal here is to develop new lead candidates for drug development through virtual screening of chemical databases to targets of known three-dimensional structure [14]. Similar techniques have recently been applied to the design of receptor and sensor proteins of novel function [15] and may be useful in the engineering of interfaces between biological and artificial nanomaterials [16].

2. Biomolecular structure prediction

2.1. Optimization methods

Stochastic optimization methods are now being used in a multitude of applications, ranging from circuit design on silicon wafers to airline flight schedules. In these and many other applications the objective is to minimize a given cost function that depends on a large number of discrete or continuous variables [11, 17, 18]. In analogy to physical problems, the cost function describes a potential energy surface (PES) in the parameter space and its global minimum optimizes the desired objective. Stochastic optimization methods are applied when enumerative methods are too costly. This is generically the case in high-dimensional optimization problems, where the total number of possible configurations grows exponentially with the number of variables.

Stochastic optimization methods successively improve one or several configurations of the underlying model to obtain an approximation of the global optimum of the PES. The optimization process thus maps onto a fictitious dynamical process of one or several configurations that move in the configuration space. The process stops when either a certain previously defined amount of computational resources have been used or when the dynamical process terminates in a stable configuration. In either case there is no guarantee that the stochastic process has found the global optimum of the PES. The computational difficulty in stochastic optimization methods depends strongly on the number of degrees of freedom and the complexity of the PES.

The simplest stochastic optimization method, repeated local optimization starting from random initial conditions, will therefore also require an exponentially large number of steps for NP-incomplete problems. To significantly reduce the computational effort, stochastic optimization methods must therefore also move uphill. In simulated annealing (SA) [8] this is achieved by simulating the finite temperature dynamics of the system. Starting from a configuration r with energy $E(r)$ one generates a new configuration r' with energy $E(r')$ which replaces the original configuration with probability:

$$P = \begin{cases} \exp(-\beta[E(r') - E(r)]) & \text{if } E(r') > E(r) \\ 1 & \text{otherwise,} \end{cases} \qquad (1)$$

where $\beta = 1/kT$ is the fictitious inverse temperature. At any given temperature such an (ergodic) Monte Carlo process [19] samples the configurations r of the PES according to their thermodynamic probability. Thus, at high temperature moves with or against the gradient are accepted with almost equal probability. At low temperature only downhill moves are accepted. In SA one thus starts with high-temperature simulation and gradually cools the system to zero temperature. If *ergodicity is not lost* during the cooling schedule, the simulation will stop in the global minimum of the PES with probability 1. However, for rugged PES, such as those encountered in PSP, SA routinely fails.

As one approach to circumvent this problem, we have recently developed the stochastic tunnelling (STUN) method [10], which incorporates the ability to escape metastable states by letting the particle in the minimization process 'tunnel' forbidden regions of the PES. As in SA we retain the idea of a biased random walk, but apply a non-linear transformation to the PES:

$$E_{\mathrm{STUN}}(x) = 1 - \exp[-\gamma(E(x) - E_0)] \qquad (2)$$

where E_0 is the lowest minimum encountered by the dynamical process so far. Alternatively a suitable upper bound for the global minimum can be used for E_0. This effective potential preserves the locations of all minima, but maps the entire energy space from E_0 to the maximum of the potential onto the interval [0, 1].

At a given finite temperature of O(1), the dynamical process can therefore pass through energy barriers of arbitrary height, while the low-energy region is resolved even better than in the original potential. The degree of steepness of the cut-off is controlled by the tunnelling parameter γ. Figure 1(b) illustrates the STUN PES for a 1D model potential (see below) at a hypothetical point in the simulation where the minimum indicated by the arrow has been found as the present best estimate for the ground state.

2.2. Biomolecular force field

Over the last decades many classical force fields [20–23] have been developed to investigate numerous phenomena in physical, organic and inorganic chemistry. The difficulties encountered in PSP justify the development of specific force fields for the following reason: their molecular building blocks, i.e. the amino acids, are well defined and limited in number. The chemical complexity associated with the design of a force field specific to peptides and proteins is therefore less than that of generic organic substances. By exploiting the fact that only a limited number of building blocks will occur, their ingredients may be specifically adapted to provide a more accurate description of the system. Also, we are interested only in the low-energy conformations of the model. As a result, many degrees of freedom that are associated with covalent interactions, e.g. bond stretching, may be neglected in the description of the system.

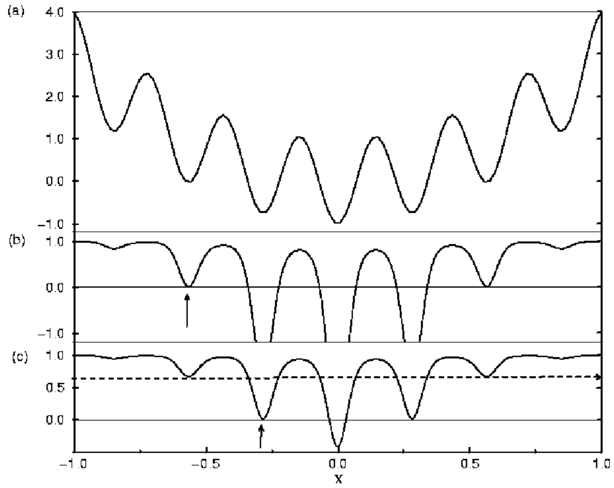

Figure 1. Schematic one-dimensional PES and its transformations under the STUN procedure, provided that the local minima indicated by the arrows have been found [10]. Part (a) shows the original PES, parts (b) and (c) the transformed PES under the assumption that the minima indicated by the arrow are the best configurations found so far in the simulation, respectively.

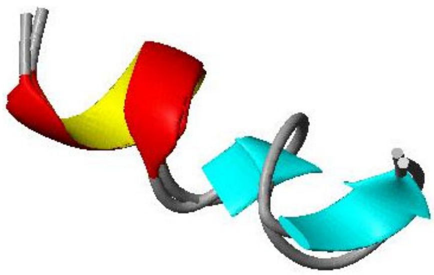

Figure 2. Overlay of the crystal structure of a 13-residue helical fragment of 1HRC (residues 92–105) with the structure obtained in the simulation.

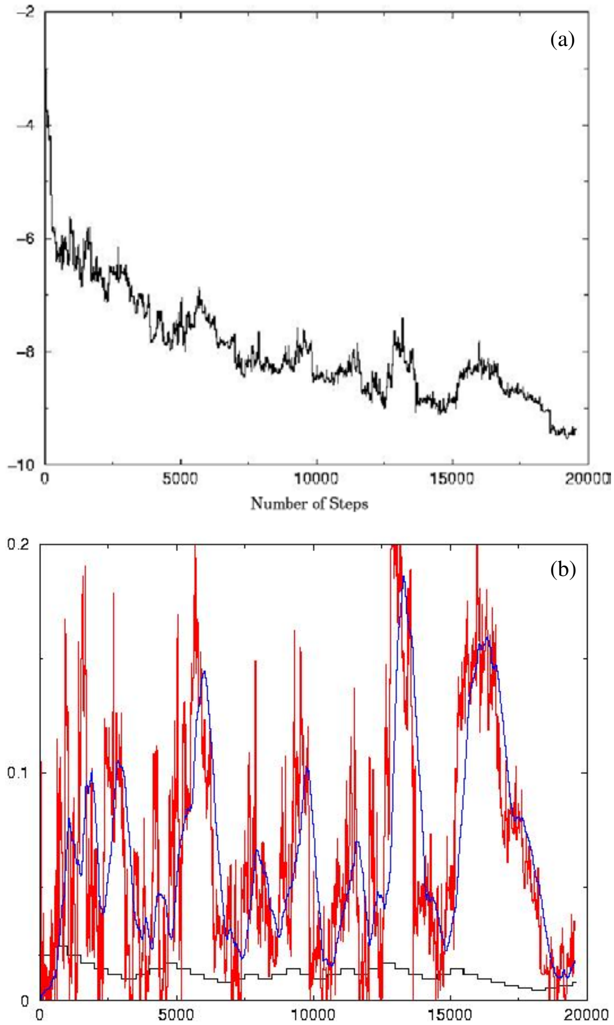

Figure 3. Application of the STUN to the folding of a 13-amino-acid helix fragment of 1HRC (residues 92–105). Part (a) shows the total energy of the system as a function of the number of simulation steps. Part (b) shows the effective energy, its moving average (dashed) and the effective temperature (bottom) of the STUN procedure. Both tunnelling and local search phases are relevant to determine the native structure of the peptide; note that tunnelling phases with relatively high effective energy correspond to large fluctuations of the original energy in (a).

The PFF01 force field [7] represents all atoms except apolar CH_n individually. CH_n groups are approximated by a single sphere comprising both the carbon and the hydrogen atoms (united atom approach). We have fitted the Lennard-Jones (LJ) radii in PFF01 to a subset of 134 proteins of the PDB database. The associated LJ interaction strength was taken from the OPLS force field [22]. We note that in simulations with explicit solvent molecules there are LJ interactions between peptide and solvent atoms. This atom-dependent effect has been incorporated into the implicit solvent model. Coulomb interactions in proteins are complicated, in particular regarding screening effects of the solvent. In the PPF01 force field we have implemented an approach which models this effect with group-dependent and interaction-dependent effective dielectric constants [24]. For the implicit solvent model, the simplest conceivable choice assigns a free energy of solvation proportional to the effective contact area each atom of the protein/peptide has with the solvent. We have subdivided the atom types of the force field into suitable subgroups and fitted the resulting model to the available experimental Gly–X–Gly data [25].

2.3. Applications

We first investigated the folding of small peptide fragments that are believed to assume a unique three-dimensional structure even when removed from their environment in the protein. Figure 2 shows the overlay of the crystal structure of a helical 13-amino-acid residue fragment of the 1HRC protein with the structure we have obtained in STUN simulations. Encouragingly, the backbone configurations of these two structures are identical to better than experimental resolution. Figure 3(a) shows the evolution of the total energy of the structure from an unfolded configuration to the folded configuration as a function of the number of energy evaluations. Figure 3(b) shows the effective energy and the effective temperature. Several heating and cooling cycles were required to fold the helix fragment, and 'tunnelling phases' that occur when the effective energy is relatively high significantly

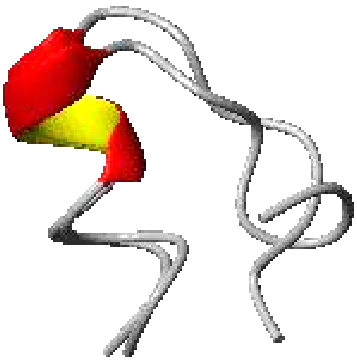

Figure 4. Overlay of the crystal structure of a helical bend in 1UBQ with the simulated structure.

aided the search process. In these phases the original energy of the system undergoes significant fluctuations that are much larger in magnitude than the difference in energy of two successive metastable states. Circumnavigating these energy barriers in a traditional simulation would significantly slow the optimization process. We conducted several dozen STUN runs for this, as well as for other fragments that were investigated, to verify that the structure we had obtained corresponds to the global optimum of the system. We noted that in SA the helix could not be folded even with a tenfold increase in computational effort. Hence STUN appears to present a viable and efficient optimization strategy to optimize peptide fragments of this length. Helical segments are stabilized by the short-range hydrogen bonds. We found that it is possible to artificially destabilize the helical structure if the prefactor of the solvent interactions is increased to unphysical values.

An example for a non-helical 12-amino-acid fragment of the 1UBQ protein is shown in figure 4. In this structure hydrogen bonding interactions that attempt to stabilize a helix compete with longer-range hydrogen bonding and solvent interactions to form a structure that is part helix part bend. The figure again illustrates the good overlap that was found in our STUN simulations for the simulated configuration and the corresponding crystal structure. A prerequisite for this success is a good balance between hydrogen bonding terms and solvent interactions in the force field.

We have also investigated the 36-residue headpiece of the villin protein that was recently simulated with molecular dynamics [23, 26]. The best configuration obtained with about a CPU week on a single PC is shown in figure 5(b) in comparison with the NMR structure. The fraction of native contacts was similar in both studies, although more than 85 years of CPU time were invested in the molecular dynamics (MD) simulation on a 256-node CRAY-T3E supercomputer. This comparison illustrates the increase in efficiency that can be obtained through the use of stochastic optimization methods, even though both simulations failed to reach the NMR structure. We find, however, that the structure obtained in our simulation has a lower (free) energy that that of the NMR structure, indicating that this failure is not due to a failure of the optimization strategy, but is attributable to a shortcoming of the force field.

This suggests a rational decoy strategy to systematically improve the force field that we will now implement. We

(a)

(b)

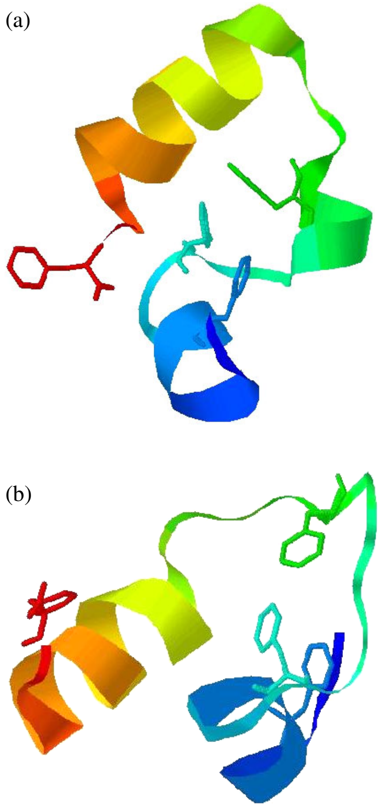

Figure 5. Comparison of the (a) NMR structure and the (b) simulated structure of 1VII.

generate a large set of 'good' candidates that compete with the NMR structure. As long as one of these decoys has a better energy than the native configuration, the force field must be modified to stabilize the native configuration in comparison to all other decoys. When this is achieved we generate new decoys by refolding the peptide, generating new configurations that are either yet again better in energy than the NMR structure or ultimately folding the peptide. In the following we report the preliminary results of this project.

We have created a set of decoys starting either with stretched configurations or the NMR configuration. Some of the latter runs were modified with an additional harmonic constraint that limited the deviation of the simulated structure from the NMR structure to 2–3 Å. The adjustable force field parameters were the surface free energies that enter the implicit solvent model, which were permitted to vary by 20% around their original values. The rationale behind this approach was that these parameters are relatively uncertain, as they are transferred from small-molecule data to very large systems. We finally arrived at a decoy set containing about 11 000 entries that each had a backbone root mean square deviation (RMSD) of at least 3 Å with every other decoy. This decoy set yields an approximate representation of the local minima of the FES of the peptide.

The main results of this analysis can be summarized as follows: (i) The lowest conformation (figure 6(N)) had a backbone deviation of only 3.5 Å from the NMR structure. The optimized force field is therefore the first to correctly predict the tertiary structure of HP36 at the all-atom level. (ii) There are only very few distinct low-lying minima of the

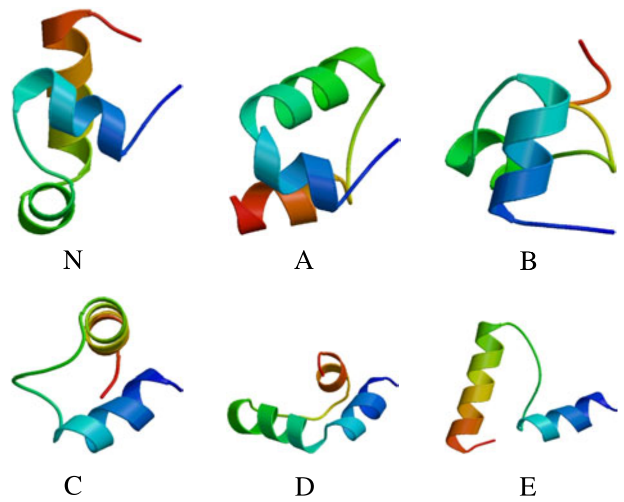

Figure 6. Representatives of the best decoy families of HP36/1VII.

FES in our model. Three of these structures (figures 6(A), (B), (D)) correspond to three-helix structures never seen in previous investigations [23, 26], while others have only two helices. Overall this picture is consistent with the existence of a complex folding funnel in the FES. From the standpoint of secondary structure analysis the low-energy structures of this funnel contain only three-helix structures. Within the folding funnel the configuration explores a subspace of the full FES in which helix length and position vary. Surprisingly there is almost no correlation in the RMSD between these structures. We note that only configurations with a deviation of more than 1 Å were counted as individual decoys. With backbone RMSDs in excess of 7 Å, the families of low-lying structures differ as much among one another as they differ with a random configuration of the decoy set.

3. Receptor–ligand docking

The goal of receptor–ligand docking is to assist the synthetic process of drug discovery through the suggestions of suitable starting candidates, so-called leads. These leads are obtained by ranking a database of candidates, which should ultimately comprise all synthesizable, non-toxic molecules, according to their affinity to a structurally resolved protein receptor. To obtain this estimate, the best position and conformation of each molecule of the database has to be determined with a suitable *docking method*. The best conformation/position of the molecule is obtained as the global optimum of a suitably defined *scoring function*. In the following we illustrate the applicability of the STUN method to this problem with two screening runs in which a methotrexate (MTX) and a set of 10 known thymidine kinase (TK) inhibitors were mixed into a database of 10 000 ligands of a database used for drug-development (nicopen). We then docked all ligands against the crystal structures of the receptor and were able to identify the known ligands on the basis of the screening run.

3.1. Force field

To test the applicability of the STUN method to receptor–ligand docking [27] we prepared a subset of the open part of the NCI database (nciopen.mol), which had been processed to generate the three-dimensional structures with the help of Corina [28, 29] (nciopen3d.mol). We identified rotatable bonds with a simple algorithm which searched for single bonds, excluding ring structures, trivial single-atom end groups and atoms with sp^2 hybrid orbitals. Among 125 000 compounds available, the first 10 000 were selected which satisfied the conditions of having not more than 100 atoms and not more than 10 rotatable bonds.

The protein coordinates were taken from the x-ray structure of *Escherichia coli* dihydrofolate reductase (DH) with MTX (pdb entry: 4dfr, monomer B) [30] and the x-ray structure of TK (pdb entry: 1kim) respectively. Hydrogen atoms and partial charges were attached using InsightII with the esff force field. The binding site was defined with the docked ligand in the x-ray structure: a volume of 5 Å radius was defined around the centre of mass position of the docked complex. If the global minimum on the PES of any ligand was located such that either the centre of mass of this ligand or more than 20 ligand atoms were inside this volume, the ligand was regarded as docked.

For the simulations discussed below we used the following scoring function:

$$S = \sum_{\text{Protein}} \sum_{\text{Ligand}} \left(\frac{R_{ij}}{r_{ij}^{12}} - \frac{A_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon r_{ij}} \right) + \sum_{\text{H-bonds}} \cos \Theta_{ij} \left(\frac{\tilde{R}_{ij}}{r_{ij}^{12}} - \frac{\tilde{A}_{ij}}{r_{ij}^{10}} \right), \quad (3)$$

which contains the empirical Pauli repulsion (first term), the van der Waals attraction (second term), the electrostatic Coulomb potential (third term) and the angular-dependent hydrogen bond potential (terms four and five) [31].

3.2. Applications

For dihydrofolate reductase 6100 ligands, out of 10 000 compounds, reached configurations where the ligand was embedded in the cavity and the external binding energy, scored with equation (3), was below -50 kJ mol^{-1} [31]. It is well known that scoring functions as the one used here are too inaccurate to yield a quantitatively approximate natural affinity of the ligand. Nevertheless they often provide a useful relative ranking of leads when docked under identical conditions. Figure 7 (left panel) shows the external binding energies of the docked compounds. Among them, the natural ligand MTX clearly scored best, with a minimum energy conformation which differed by only 1.4 Å from the x-ray crystal structure (figure 8). This figure shows that both conformations essentially differ in the positions of only one of their carboxylate groups. In the minimal conformation of the scoring function this group is turned around to create an extra hydrogen bond to Lys-32 which is absent in the crystallographic conformation. In the experimental configuration a conserved water molecule (Wat-672) mediates competing hydrogen bonds in the natural environment, an effect which cannot be accounted for with the present scoring function.

For the TK inhibitors seven of ten docked in the receptor, but only three inhibitors were ranked with a high affinity (see table 1). The degree of database enrichment, i.e. the fraction

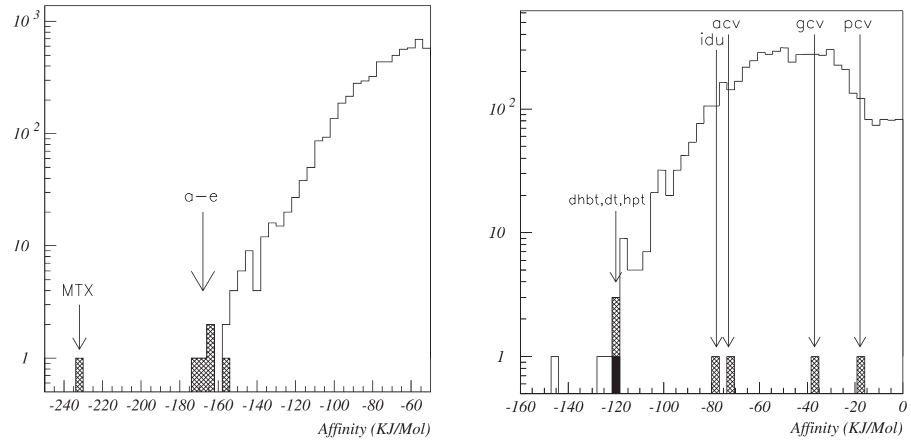

Figure 7. Histogram of the affinities of 10 000 ligands docked to dihydrofolate reductase (left) (see [31]) and thymidine kinase (TK) (right). For the 4dfr docking run, the natural ligand (MTX) was scoring best by a wide margin, for TK, 7 of 10 known inhibitors docked successfully. The inhibitors associated with the labels shown in the figure are shown in [32].

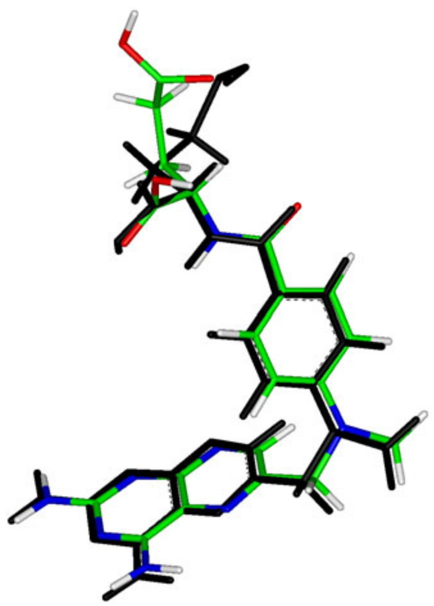

Figure 8. MTX in its crystallographic x-ray conformation (black) and the global minimum found with STUN (grey). Both conformations differ in the positions of their carboxylate groups.

of well-scoring ligands versus the total number of ligands, is thus comparable to other recent studies of the same model system [32]. We have verified, by repetition of the docking runs, that the affinity of the individual inhibitors is accurately computed in the model, deficiencies in the ranking of the known inhibitors thus result from inadequacies of the scoring function.

4. Summary and conclusions

We have proposed the use of stochastic optimization methods to predict the structure of complicated biomolecules. To implement this approach, we have developed STUN as a suitable optimization algorithm to deal with the very rugged PESs that often occur in biomolecular/nanomaterial simulations.

Table 1. Value of the scoring function (affinity) and rank of the 10 thymidine kinase inhibitors in a scoring run against 10 000 molecules from the nciopen database. The structures associated with the inhibitors are given in [32].

Inhibitor	Affinity	Rank
dhbt	−120.584	4
dt	−120.422	5
hpt	−119.079	6
idu	−76.918	515
acv	−72.550	719
gcv	−36.632	3351
pcv	−18.970	4845
ahiu	NA	NA
hmtt	NA	NA
mct	NA	NA

We developed a biomolecular force field (PFF01) that parametrizes the free energy of the underlying system with an implicit representation of the interactions of the biomolecule with the solvent. We have argued that there is a rational, decoy-based strategy to adapt such force fields that can be used to predict the structure of peptides and proteins using stochastic optimization techniques such as the STUN. We have illustrated this approach in the folding of short peptide fragments and presented an analysis of the difficulties encountered in the folding of the 36 head residues of 1VII.

Secondly we have given an overview of an automated virtual screening of 10 000 compounds to the active site of *Escherichia coli* dihydrofolate reductase (4dfr) and to the active site of TK. The global optimization technique STUN, which has already been successful in the rigid ligand approximation, also proved reliable and efficient for the more demanding application to flexible ligands. It was able to match the natural ligand MTX (which contains nine rotatable bonds) to an accuracy of 1.4 Å with the x-ray structure. The scoring function ranked this ligand as the best among the 10 000 screened molecules, validating both the optimization method and scoring function. In a similar screen, three of the ten known inhibitors of TK were correctly identified among the best-scoring ligands in the screening procedure, while three other ligands docked, but were badly ranked. The failure to correctly estimate their binding affinity could be traced to deficiencies in the scoring function (force field) rather than to the optimization method.

These applications illustrate both the success and the limitations of the use of stochastic optimization methods for the *de novo* simulation of biomolecular nanomaterials. Both free-energy force fields and simulation/optimization techniques must be continuously developed to be able to contribute to our understanding of the often complex mechanisms governing the behaviours of these systems.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (We 1863/11-1), the BMWF and the Bode foundation. We are grateful to J Moult and S Gregurick for helpful discussions.

References

[1] Baker D and Sali A 2001 *Science* **294** 93
[2] Davis A P 1999 *Nature* **401** 120
[3] Schliewa G W M 2001 *Nature* **411** 424
[4] Soong R K, Bachard G D, Neves H P, Olkhovet A G, Craighead H G and Montemagno C D 2000 *Science* **290** 1555
[5] Goddard W A III 2000 *Bionanotechnology—de novo Simulations and Design* http://www.foresight.org/Conferences/MNT8/Abstracts/Goddard/index.html
[6] Robson B 1999 *Trends Biotechnol.* **17** 311
[7] Herges T, Merlitz H and Wenzel W 2002 *J. Ass. Lab. Autom.* **7** 98
[8] Kirkpatrick S, Gelatt C and Vecchi M 1983 *Science* **220** 671
[9] Hansmann U and Okamoto Y 1999 *J. Phys. Chem.* B **103** 1595
[10] Wenzel W and Hamacher K 1999 *Phys. Rev. Lett.* **82** 3003
[11] Wales D and Scheraga H 1999 *Science* **285** 1368
[12] Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucleic Acids Res.* **28** 235 http://www.rcsb.org/pdb
[13] Schonbrunn J, Wedemeyer W J and Baker D 2002 *Curr. Opin. Struct. Biol.* **12** 348
[14] Walters W, Stahl M and Murcko M 1998 *Drug Discovery Today* **3** 160
[15] Looger L L, Dwyer M A, Smith J J and Hellinga H W 2003 *Nature* **423** 185
[16] Liu H, Schmidt J J, Kizk G D, Bachand S S, Looger L L, Hellinga H W and Montemagno C D 2002 *Nature Mater.* **1** 173
[17] Brooks C L, Onuchic J N and Wales D J 2001 *Science* **293** 612
[18] Piela L, Kostrowicki J and Scheraga H A 1989 *J. Phys. Chem.* **93** 3339
[19] Metropolis N, Rosenbluth A W, Augusta M N R, Teller H and Teller E 1953 *J. Chem. Phys.* **21** 1087
[20] van Gunsteren W and Berendsen H 1987 The Groningen molecular simulation manual (GROMOS) *Technical Report* Groningen University
[21] MacKerell A D Jr *et al* 1998 *J. Phys. Chem.* B **102** 3586
[22] Jorgensen W L and McDonald N A 1998 *J. Mol. Struct.* **424** 145
[23] Duan Y and Kollman P A 1998 *Science* **23** 740
[24] Avbelj F and Moult J 1995 *Biochemistry* **34** 755
[25] Eisenberg D and McLachlan A D 1986 *Nature* **319** 199
[26] Hansmann H 2002 *Phys. Rev. Lett.* **88** 068105
[27] Merlitz H and Wenzel W 2002 *Chem. Phys. Lett.* **362** 271
[28] Sadowski J and Gasteiger J 1993 *Chem. Rev.* **93** 2567
[29] Sadowski J, Gasteiger J and Klebe G 1994 *J. Chem. Inf. Comput. Sci.* **34** 1000
[30] Bolin J, Filman D, Matthews A and Kraut R H J 1982 *J. Biol. Chem.* **257** 13650
[31] Merlitz H, Burghardt B and Wenzel W 2003 *Chem. Phys. Lett.* **370** 68
[32] Bissantz C, Folkerts G and Rognan D 2000 *J. Med. Chem.* **43** 4759