Benchmarking the POEM@HOME Network for Protein Structure Prediction

Timo Strunk[1], Priya Anand[1], Martin Brieg[2], Moritz Wolf[1], Konstantin Klenin[2], Irene Meliciani[1], Frank Tristram[1], Ivan Kondov[2] and Wolfgang Wenzel[1,*]

[1]Institute of Nanotechnology, Karlsruhe Institute of Technology, PO Box 3640, 76021 Karlsruhe, Germany.

[2]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, PO Box 3640, 76021 Karlsruhe, Germany

ABSTRACT

Motivation: Structure based methods for drug design offer great potential for in-silico discovery of novel drugs but require accurate models of the target protein. Because many proteins, in particular transmembrane proteins, are difficult to characterize experimentally, methods of protein structure prediction are required to close the gap between sequence and structure information. Established methods for protein structure prediction work well only for targets of high homology to known proteins, while biophysics based simulation methods are restricted to small systems and require enormous computational resources.

Results: Here we investigate the performance of a world-wide distributed computing network, POEM@HOME, which implements a biophysical model for protein modeling, as a robust computational infrastructure for protein structure prediction. We demonstrate the use of this network for the time-consuming energy relaxations for decoy sets and two targets of the 2010 protein structure prediction assessment (CASP).

Conclusion: We demonstrated the use of the POEM@HOME network as a robust computational resource for protein structure prediction based on relaxation in biophysical models. Efforts to implement a web-interface to make this resource available to life-science researchers are presently under way.

1 INTRODUCTION

With the completion of sequencing efforts for many important genomes, protein structure and function prediction emerges as an important goal to make progress in structure based drug design (Kryshtafovych, et al., 2007; Moult, et al., 2005). Methods for protein modeling have a wide variety of objectives, such as structure prediction, molecular replacement, prediction of protein stability/disorder or property prediction of mutations. Physics-based or forcefield-based methods, which were initially believed to hold great promise for protein structure prediction, now play only a marginal role in the (participant blind) biannual comparative assessment of methods for protein structure prediction (CASP) (Kryshtafovych, et al., 2005). Presently, most models submitted to this computational experiment originate from bioinformatics based methods (Kryshtafovych, et al., 2007). One reason for this state-of-affairs is the high computational cost of all-atom forcefield-based models. However, even for computationally feasible problems, for example for structure refinement (Das, et al., 2007), recent investigations point to deficiencies for most of the presently available forcefields (Fitzgerald, et al., 2007). Knowledge-based potentials, in contrast, perform very well in differentiating native from non-native protein structures (Wang, et al., 2004; Zhou, et al., 2007; Zhou, et al., 2006) and have recently made inroads into the area of protein folding. Physics-based models retain the appeal of high transferability, but the present lack of truly transferable potentials calls for the development of novel forcefields for protein structure prediction and modeling (Schug, et al., 2006; Verma, et al., 2007; Verma and Wenzel, 2009).

We have earlier reported the rational development of transferable free energy forcefields PFF01/02 (Schug, et al., 2005; Verma and Wenzel, 2009) that correctly predict the native conformation of more than 27 small proteins in simulations starting from a completely extended structure. In order to perform these simulations we have developed an increasingly sophisticated set of sampling methods of the low-energy landscape of the system (Herges, et al., 2004; Schug, et al., 2005; Schug and Wenzel, 2004). Because the computational effort of the simulations increases very rapidly with system size, simulations for large systems are only feasible if a large number of processors can be exploited. In contrast to kinetics based simulation approaches, such as molecular dynamics, our approach permits splitting the simulations into several independent tasks (Verma, et al., 2007; Verma, et al., 2008). We have experimented with a number of such schemes and found evolutionary algorithms, which evolve a population of conformations in a coarse grained parallel fashion, to be very effective. Using PC clusters and high-performance computational architectures we were able to fold small proteins with up to 60 amino acids using tens of thousands of short independent simulations. Analyzing these simulations we noted that the inherent parallelism of the protocols is so large that we might as well use grid computing (or cloud computing) resources to perform the simulations.

We therefore implemented our algorithm in a world-wide volunteer computational network, POEM@HOME, which has been operational since 2007 and has grown to over 60.000 participants in more than 100 countries, delivering an average performance of over 20TFLOP/s in 2010. While such a network delivers a significant computational power, it is clearly unsuited for inherently sequential simulations. In this investigation we therefore wanted to test its performance for protein structure prediction (Gopal, et al., 2009) as part of an ongoing effort to provide the life-science community with a POEM based protein structure prediction server. Here we therefore report the overall characteristics of the POEM@HOME network and results obtained in two characteristic applications for the development of methods for protein structure prediction. In the first application we used POEM@HOME as a workhorse for ranking large decoys sets to validate the selectivity of the underlying forcefield PFF02. It is well known that present-day forcefields are not of sufficient accuracy to deliver protein

*To whom correspondence should be addressed: wolfgang.wenzel@kit.edu

structure predictions with experimental resolution due to inherent forcefield errors. To improve these models requires very large scale computations in which the ranking of near-native conformations in large decoy sets of competing structures is monitored for many proteins as a function of the forcefield parameters. In the second application we report the performance of POEM@HOME for structure prediction for two targets of last year's competitive assessment of methods for protein structure prediction (CASP). In this exercise many computational groups compete to blindly predict the experimentally known, but not yet released, structure of proteins. In this investigation we report two complementary experiments, for T0537 and T0643, a high- and low-homology target in this competition, respectively. Because predictions in CASP must be returned within three weeks of target release, use of a BOINC based network with very long average return times poses a significant challenge.

2 METHODS

2.1 Forcefield

All-atom refinement: POEM (Protein Optimization using Energy Methods) is an all-atom free-energy protein simulation package implementing the free-energy model PFF02 (Verma and Wenzel). PFF02 models the relevant protein interaction energy terms through five semi-empirical terms. The attractive and repulsive van-der-Waals forces are modeled using a 6-12 Lennard Jones potential. Electrostatic interactions could be described via a simple 1/r vacuum potential modified by the exposed surface area of the interacting groups. An implicit solvent model is employed to represent the protein-solvent interaction. The exposed surface area of each atom is multiplied by a hydrophobicity index and then accumulated. Hydrogen bonds are described via dipole-dipole interactions included in the electrostatic terms and an additional short-range term for backbone-backbone hydrogen bonding. In addition to the terms already present in PFF01, the forcefield PFF02 contains an additional term, i.e. a torsional potential for backbone dihedral angles. This force field was demonstrated to select near-native decoys for all 32 monomeric proteins (without cofactors) from the ROSETTA decoy set (Tsai, et al.) and used to fold a set of 24 proteins with helical, sheet and mixed secondary structure in de novo simulations (Verma and Wenzel).

2.2 Relaxation Protocol

Protein structures in this study were relaxed in the PFF02 forcefield to allow the unbiased comparison of structures constructed from different sources. Single relaxation simulations consist of a fixed number of Monte Carlo steps changing main- and side-chain dihedral angles of the simulated protein by a random angle. In case of proteins with several chains, also center-of-mass degrees of freedom between the different chains are changed in the simulation. After each Monte Carlo step the Metropolis criterion is evaluated and the new conformation either rejected or accepted to achieve detailed balance. During the simulations structures are annealed using a geometrical temperature scaling scheme. The protein's high-dimensional conformational space necessitates parallel sampling, which can be achieved by starting relaxation simulations in various directions from an initial structure. Therefore a multitude of single simulations were run for each initial structure. The conformation with the lowest energy was then used as the final prediction.

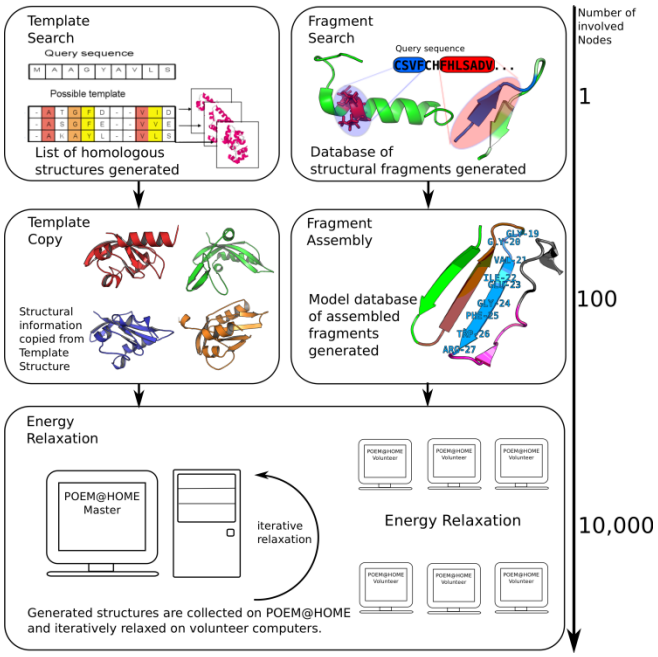

Fig 1: Schematic of the prediction protocol: Two parallel workflow branches predict initial models using homology modeling (left) and heuristic fragment assembly (right). The homology modeling workflow (left) searches for similar sequences among the database of all known experimental structures. Structural information from these models is then used to build structure candidates. Small parts of the sequence are matched using a fragment database of known structural segments. These are then assembled to full models of the whole protein. Models generated using these two branches are accumulated and relaxed on the POEM@HOME volunteer architecture. The best energy structure is chosen as the final prediction.

3 PROTEIN STRUCTURE SIMULATION

In the following we report on the use of the POEM@HOME world-wide distributed volunteer network for protein structure prediction in the context of two targets of the CASP9 protein structure prediction exercise. The general prediction protocol is summarized in Fig 1. Given the target we first search for homologous proteins for which an experimental structure is known. If this is the case (high homology target), we identify all such templates and generate initial models which then need to be ranked in energy. If no homologous targets are known (low-homology targets), we use heuristic methods to generate a large set of possible conformations, which are then again ranked in energy in our forcefield. Because the energy landscape is very rough, ranking the starting models generates very noisy predictions. For this reason we need to perform a short relaxation simulation, which attempts to map each model to a nearby local minimum (see methods section). In order to demonstrate the success of this approach for a system where the result was known, we precede the examples from CASP with the analysis of the ranking of published decoy sets for two test proteins, where the experimental structure is already known.

3.1 POEM@HOME

POEM@HOME is a distributed volunteer computing architecture implemented using the BOINC (Anderson 2004) framework. A BOINC server holds a database of workunits, which are scheduled to run on computers of volunteers, participants of the project, in remote locations. The BOINC client decides when to download new work units, when to compute them and when to return the results, however the user has options to constrain runtime and time of day for the simulations. This imposes several constraints on the types of work units that can be processed as well as on the type of algorithms that can be used. Single workunits should not exceed four hours in runtime and one Megabyte in space, as otherwise either common DSL connections are inadequate for transferal or PCs are simply shut off. Furthermore job processing has to be asynchronous as work units cannot be expected to return in time. Asynchronous means that jobs sent at the same time return at different times due to the BOINC scheduling and the users' settings in the BOINC client. Lost jobs are rescheduled automatically; a work unit can however never be guaranteed to return. Figure 2, top plot, shows the turnaround time for a work unit with an average 1 hour compute time demonstrating the asynchronous behavior of the sent workunits. Independent from the compute time of the

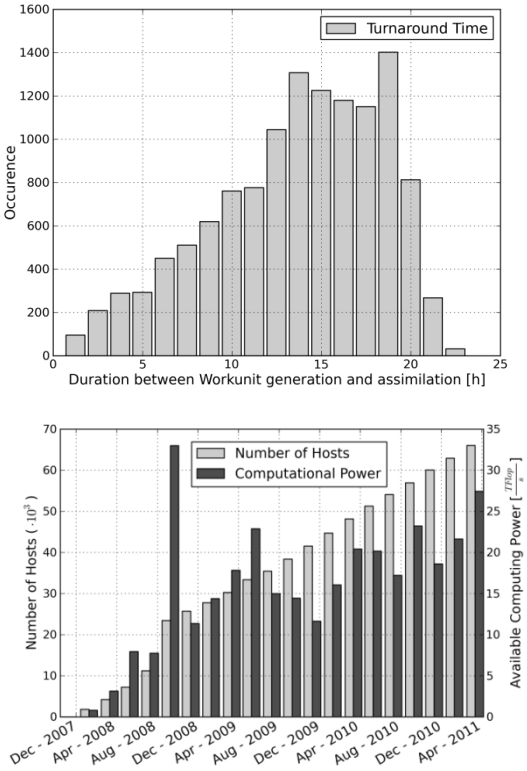

Fig 2: Top**:** Histogram of return times of one batch of 13000 relaxation jobs submitted at the same time. Assimilation means the moment in time, when the BOINC server registers the arrival of the completed workunit; Bottom**:** Growth of the computational power of the POEM@HOME network as a function of time. The graph shows both the growth in users and in the computing power. The peak in computational time during September 2008 is related to a local competition on our server.

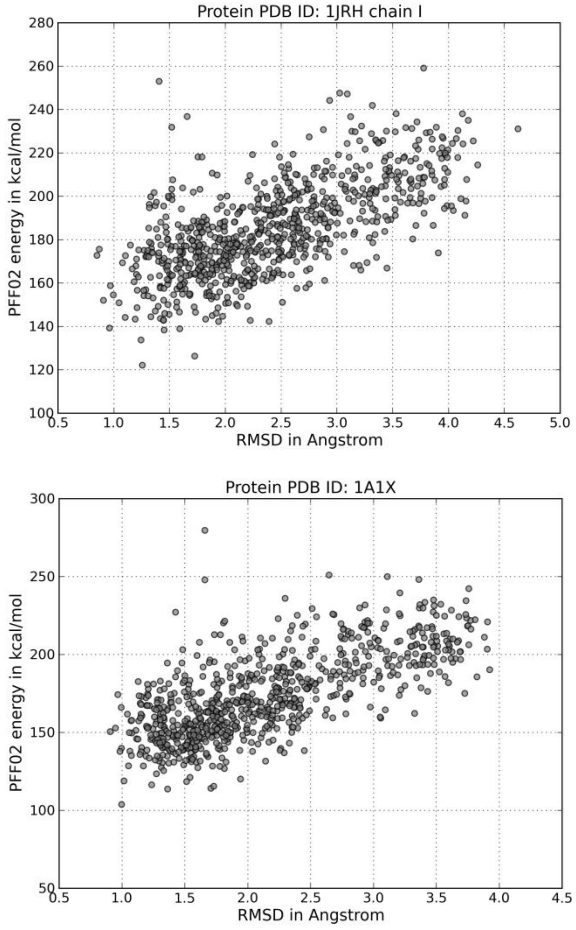

Fig 3: Scatter plots of PFF02 energies and root mean square deviations to native structure (RMSD) for proteins 1JRH chain I (top) and 1A1X (bottom) from the decoy set. Both plots show a correlation between the RMSD and the simulated energies.

work unit, the graph presents an expected turnaround time of 15 hours.

POEM@HOME runs on two machines, a MySQL and BOINC daemon host with 8 Intel Xeon 5130 cpus with 16 GB RAM and 160GB of host memory on 10,000 rpm disks in a RAID 1 configuration and on a storage backend with 3TB of host memory on 7,200 SATA disks in a RAID 5 configuration plus hot-spare. Both are connected using Gigabit Ethernet. BOINC projects need customized validator and assimilator daemons tailored to the project. After workunit completion a client delivers a finished simulated structure and an energy fitting this structure. Validation of the structure is hence possible by simply recalculating the energy on the server once more. The assimilator then just moves the structure into an appropriate directory on the server where statistics of all simulated structures are accumulated.

3.2 Performance of POEM@HOME for decoy sets

Since relaxation and rating of decoys is the most computationally demanding task, establishing confidence in the results produced by this process is crucial. Decoy sets of proteins are generally used to analyze the selectivity of a forcefield to find near-native structures among a set of misfolded structures. Using a set containing decoys for 1400 proteins (Rajgaria, et al., 2006) we assessed the performance of the PFF02 forcefield by calculating the PFF02 energy for two exemplary proteins, 1A1X consisting mainly of alpha helices and 1JRH chain I consisting of beta sheets to demonstrate this selectivity. Plotting the structures' energies against their root mean square deviation (RMSD) of the atom positions to the corresponding native structure we can measure the ability of our force field to find near native protein structures in a set of misfolded structures (Fig 3). For both proteins we find that the RMSD of the lowest energy structure is close to the best RMSD structure. Considering the high number of degrees of freedom that proteins possess, there is also a good correlation between PFF02 energy and RMSD with correlation coefficients of 0.70 and 0.66 for 1A1X and 1JRH chain I respectively. This emphasizes the good selectivity of our force field. In 65% cases the lowest energy structure has a RMSD lower than 2.0 Å and this percentage increases to 94% for 3.0 Å (Fig. 3, bottom plot). Noting that only 37% of the proteins possess a misfolded structure with a RMSD lower than 1.0 Å, this demonstrates the good selectivity of our forcefield.

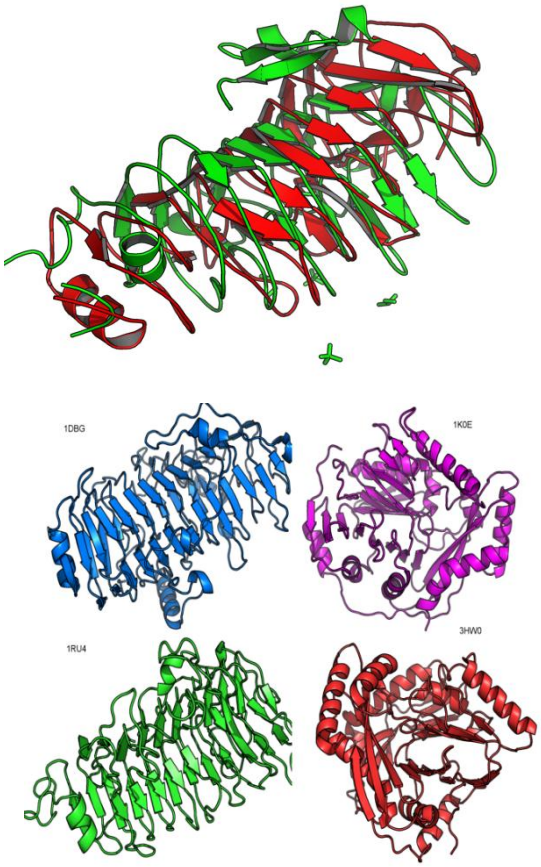

Fig 4: Top: overlay of the lowest energy model for T0537 and the corresponding experimental structure; Bottom: four possible models for T0537 based on different alignments (labeled by template protein).

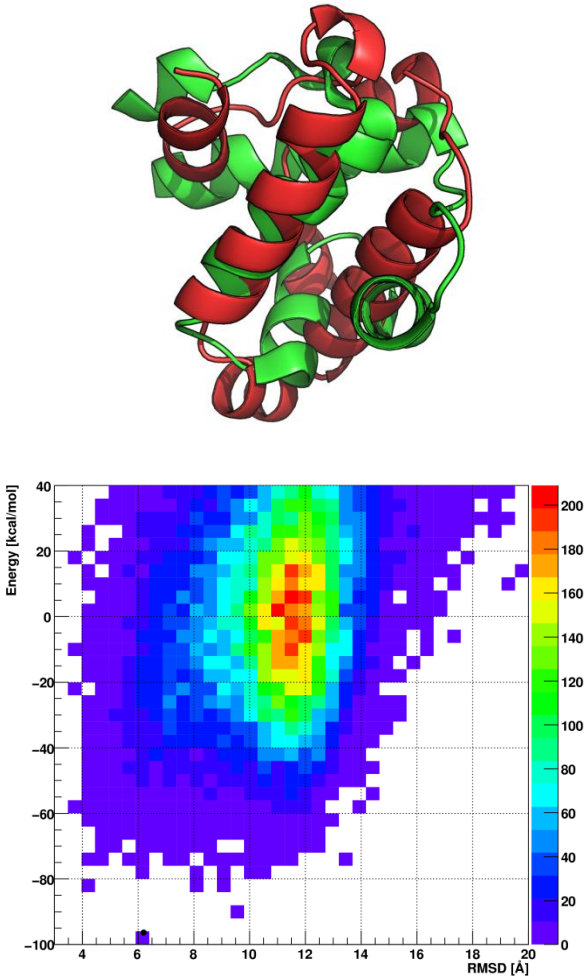

Fig 5: Top: overlay of the final result with the native conformation for the low homology target T0643; Bottom: Energy vs. RMSD plot. The black dot marks the best energy structure. A favorable energy was found for a structure with 4 Å

3.3 Performance of POEM@HOME for a high-homology target of the CASP10 evaluation

Sequence-profile alignment tools such as PSI-BLAST, 3DJury, PHYRE were used to search the 3D protein structural database PDB (the protein data bank) for homologous templates. At least one template is chosen with an alignment that covers more than 70% of the target sequence with an E-Value of $1 \cdot 10^{-3}$ or less. The E-value marks the probability that a found sequence was selected only by chance and not by apparent homology. This homologous template with high confidence alignment is then selected and a sequence alignment is generated using the clustalw program, which is used to obtain a three dimensional structure using the homology modeling protocol of the MOE program. However, if multiple templates were found with the required confidence levels, multiple structures with sufficient conformational variability were selected and modeled using MOE. One exemplary protein structure, where this modeling protocol was applied was T0537. We show the prediction of this protein due to the high homology to other proteins in the PDB database. Possible template structures for this model were 1K0E (pink), 3HW0 (red), 1RU4 (green) and 1DBG (blue) as shown in Fig. 4 (bottom). An alignment was generated for all the four templates and the alignment between the target and 1K0E and 3HW0 resulted in an overall realistic global dimer-like fold, with a beta sheet core isolated circularly by helices as shown in Fig. 4 (top). On the other hand, 1DBG and 1RU4 resulted in a completely different global fold, a beta-sheet-only tube. Energy relaxation for both all the models were done using POEM@HOME selected the 1DBG model as the best-energy model by a wide margin (~40 kcal/mol difference), which corresponded to the correct global fold. Even though human inspection favored the 1DBG homology model, because the gene-family of T0537 and 1DBG matched, leaving us undecided which model to choose. The relative RMSD between the model submitted and experimental structure is around 3.5 Å. The aligned structures are shown below in Fig. 4 (top graphic).

3.4 Performance of POEM@HOME for a low-homology target of the CASP10 evaluation

CASP target T0643 showed no apparent homology with known structures at the time of CASP. It is therefore an example for the application of our free-modeling protocol (the right branch in Fig. 1). The Rosetta 3.1 software suite was used to generate 31,000 structure proposals from a fragment database containing 16,000 fragments of length three and 15,000 fragments of length nine. These predictions took roughly 12 hours on 40 cores of AMD Opteron processors 2376. The generated decoys were assembled in the default Rosetta 3.1 prediction protocol. Afterwards they were annealed from 300 K to 5 K on POEM@HOME in 200,000 step relaxation runs. The mainchain and sidechain dihedral angles are selected for moves in a ratio of 7:3 mainchain to sidechain.

Figure 5, bottom plot, shows energies and RMSDs of all generated structural models for target T0643. The best energy structure features an energy of −95 kcal/mol and a RMSD of 6 Å to the experimental structure. Of the five submitted structures, the best structure in comparison to the native one has an RMSD of 4 Å with an energy of −85 kcal/mol.

4 DISCUSSION

Biophysics-based methods for protein structure prediction are significantly more demanding computationally than their counterparts using heuristic scoring functions. However, recent progress in the development in force fields and simulation methodology increasingly places biophysics based modeling techniques for protein structure prediction within reach. In order to offer such services for a wide community of life-science researchers at low/no cost substantial computational resources to perform the required simulations must be provided. In this investigation we reported the use of the world-wide distributed volunteer computation network, POEM@HOME, for protein structure prediction. We demonstrated that a decoy ranking procedure can be efficiently implemented on such a network for accurate protein structure prediction for selected targets of the last CASP exercise as well as in a decoy ranking studies. The long turnaround time (compared to the computational cost of a single work unit) makes such networks not usable for all kinds of simulations. However, for the application at hand, such delays can be tolerated even for protocols which require several relaxation iterations. We therefore conclude that such computational networks, which are also used in Rosetta@home (Bonneau, et al., 2001) or Folding@home (Snow, et al., 2004), can make a significant contribution to provide low-cost approaches to protein structure prediction. Efforts to make our biophysics based schemes available to a wide community of users via a web-interface are presently underway. We also note in closing, that it is quite easy to use other backends, such as grid- or cloud-based resources, for this type of application.

ACKNOWLEDGEMENTS

We are very grateful for the continued support of the volunteers of the POEM@HOME network (**http://boinc.fzk.de**) without whom this work would not have been possible.

Funding: This investigation has been supported by the grant for life sciences HPC-5 within the HPC program of the Baden-Wuerttemberg Stiftung.

REFERENCES

Anderson, D.P. (2004) BOINC: A system for public-resource computing and storage, *Fifth Ieee/Acm International Workshop on Grid Computing, Proceedings*, 4-10, 469.

Bonneau, R., *et al.* (2001) Rosetta in CASP4: progress in ab-initio protein structure prediction, *Proteins*, **45**, 119-126.

Das, R., *et al.* (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home, *Proteins*, **69 Suppl 8**, 118-128.

Fitzgerald, J.E., *et al.* (2007) Reduced C(beta) statistical potentials can outperform all-atom potentials in decoy identification, *Protein Sci*, **16**, 2123-2139.

Gopal, S.M., Klenin, K. and Wenzel, W. (2009) Template-free protein structure prediction and quality assessment with an all-atom free-energy model, *Proteins-Structure Function and Bioinformatics*, **77**, 330-341.

Herges, T., Schug, A. and Wenzel, W. (2004) Exploration of the free-energy surface of a three-helix peptide with Stochastic optimization methods, *International Journal of Quantum Chemistry*, **99**, 854-863.

Kryshtafovych, A., Fidelis, K. and Moult, J. (2007) Progress from CASP6 to CASP7, *Proteins*, **69 Suppl 8**, 194-207.

Kryshtafovych, A., *et al.* (2005) Progress over the first decade of CASP

experiments, *Proteins: Structure, Function and Bioinformatics*, (in press).
Moult, J., *et al.* (2005) Critical assessment of methods of protein structure prediction (CASP) - Round 6, *Proteins: Structure, Function, and Bioinformatics*, **61**, 3-7.
Rajgaria, R., McAllister, S. and Floudas, C. (2006) A novel high resolution C-alpha-C-alpha distance dependent force field based on a high quality decoy set, *Proteins-Structure Function and Bioinformatics*, 726-741.
Schug, A., *et al.* (2005) Comparison of stochastic optimization methods for all-atom folding of the Trp-Cage protein, *ChemPhysChem*, **6**, 2640-2646.
Schug, A., *et al.* (2006) Stochastic optimization methods for protein folding. In Julien, J.P., *et al.* (eds), *Recent Advances in the Theory of Chemical and Physical Systems*. pp. 557-572.
Schug, A., Herges, T. and Wenzel, W. (2005) All atom protein folding with stochastic optimization methods, *Biophysical journal*, **88**, 332a-332a.
Schug, A. and Wenzel, W. (2004) Predictive in-silico all-atom folding of a four helix protein with a free-energy model, *J. Am. Chem. Soc.*, **126**, 16736-16737.
Snow, C.D., *et al.* (2004) Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy, *Proc. Nat. Acad. Sci. (USA)*, **101**, 4077-4082.
Tsai, J., *et al.* (2003) An improved protein decoy set for testing energy functions for protein structure prediction, *Proteins*, **53**, 76-87.
Verma, A., *et al.* (2007) All-atom de novo protein folding with a scalable evolutionary algorithm, *Journal of computational chemistry*, **28**, 2552-2558.
Verma, A., *et al.* (2008) Massively Parallel All Atom Protein Folding in a Single Day, *Parallel Computing: Architectures, Algorithms and Applications*, **15**, 527-534.
Verma, A., *et al.* (2007) All atom protein folding with massively parallel computers. In Long, C.A. and Anninos, P. (eds), *Bio'07: Proceedings of the 3rd Wseas International Conference on Cellular and Molecular Biology, Biophysics and Bioengineering*. pp. 121-125.
Verma, A. and Wenzel, W. (2009) A Free-Energy Approach for All-Atom Protein Simulation, *Biophysical journal*, **96**, 3483-3494.
Wang, K., *et al.* (2004) Improved protein structure selection using decoy-dependent discriminatory functions, *BMC structural biology*, **4**, 8.
Zhou, H., *et al.* (2007) Analysis of TASSER-based CASP7 protein structure prediction results, *Proteins*, **69 Suppl 8**, 90-97.
Zhou, Y., *et al.* (2006) What is a desirable statistical energy function for proteins and how can it be obtained?, *Cell biochemistry and biophysics*, **46**, 165-174.