Modeling of structure, folding and interactions of biomolecules in the era of GPGPU computing

S. Hayryan[1], M.-C. Wu[1], C.-K. Hu[1], Z. Gažová[2] and T. Kožár[2]

[1]Insitute of Physics, Academia Sinica, Taipei, Taiwan
[2]Department of Biophysics, Institute of Experimental Physics, Slovak Academy of Sciences, Košice, Slovakia
tibor@saske.sk

Abstract. The recent boom in general-purpose computing on graphics processing units (GPGPU) facilitates simulations with high demands on computer resources. Such simulations are typical for macromolecules and nanoparticles of biological importance. Several proteins, instead of folding into biologically active 3D structures, aggregate together forming large fibril structures called amyloid aggregates. Amyloids are being extensively studied both experimentally and through computer simulations. Since amyloid aggregates are huge molecular complexes composed from hundreds of thousands of atoms, it is clear that their simulations need supercomputing power. GPGPU-based clusters were shown to offer alternative resources for performing molecular dynamics simulations on nanoscale. We were also using one of the newest docking methodology (the AutoDock Vina program) to model the differences in ligand binding to the native insulin and to the unfolded complexes. In addition, virtual lectin arrays were constructed and high-throughput *"In Silico"* screening was performed in order to select the best binders to the particular galectins.

Keywords: GPGPU computing, molecular modeling; protein structure and interactions, amyloid aggregation.

1 Introduction

A large variety of computational methods is available to calculate structural and electronic properties of biomolecules and their complexes. High demand for computational resources is common for almost all first-principle quantum mechanical methods in dependence on the level of study. This can be started from the relatively fast semiempirical methods to the *ab initio* level of solving of the Schrodinger equation (either using Hartree/Fock (HF) or Density Functional Theory (DFT)) to the full configuration interaction (CI) protocol with Moller-Plesset (MP) perturbation theory in between. The size of the molecules (the number of atoms and the appropriate selection of the number and type of base functions describing the atomic orbitals) is a key factor influencing the approximation level of the method used in the computational study of molecular properties and behavior. The computation time of the *ab initio* method scales usually with n^4 where n is the number of the atomic base functions. This scal-

ing becomes even worse for MP2 (i.e. n^5), MP4 (i.e. n^5) or coupled-cluster CI (i.e. n^7). Accordingly, such scaling reduces the usability of coupled-cluster CI to molecules of moderate size.

Biological macromolecules such as proteins or nucleic acids and their aggregates (see Figure 1) are composed from thousands of atoms and their computational studies need different approaches. Approximative methods of the molecular mechanics (MM) (or sometimes called force field (FF)) are often used to calculate the molecular energies. The typical energy function here is a sum of terms accounting for bond deformations, bond angle distortions, torsion potentials, electrostatic interactions, nonbonded interactions (e.g. Lennard-Jones potential) and hydrogen bonding.

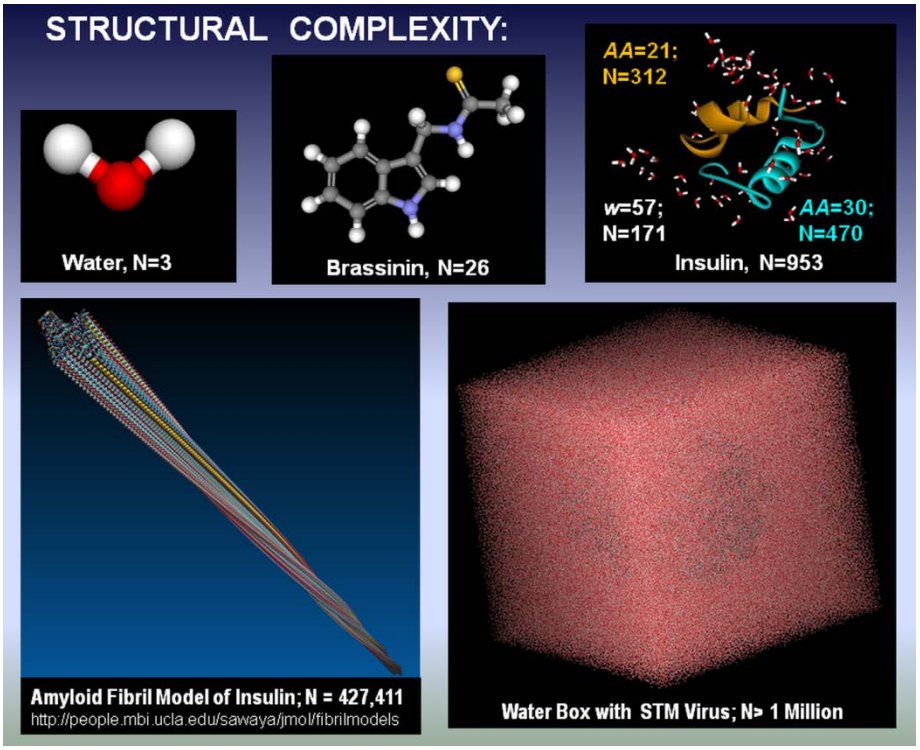

Fig. 1. Molecular examples from a three-atomic water molecules to the Satelite Tobacco Mosaic Virus (hidden in a water box [1]). Atomic color coding: Oxygen – red, Hydrogen – white, Nitrogen – Blue, Sulphur – Yellow.

Due to their simplicity the MM methods are relatively fast (scale with N^3 where N is the number of atoms in the system). However, they are not always suitable for studies of chemical reactions (bond breaking/creation of new bonds). Combined approaches of QC/MM were proposed in order to study such cases of biological interest, i.e. reactions in the enzyme active site. In QC/MM protocols the atoms (the bound ligand and the surrounding atoms of the protein) are studied on HF or DFT level while the rest of the protein is described using MM protocols.

Monte Carlo (MC) or Molecular Dynamics (MD) are utilized when the time development of the system is in question. The Newton's equation of motion is solved in MD with femtosecond integration time step. Relatively long simulation times are required for simulations of solvated biological macromolecules and their aggregates in order to gain results comparable to experimental measurements. Accordingly, such simulations have high demands on computer resources (number and speed or processors, disk storage and inter-node communication).

Approximative methodologies have been proposed also for the computations of protein-ligand or protein-protein intermolecular interaction energy profiles. These methods are described as molecular docking and the interaction energies are estimated from empirically parametrized functions. The computational demand comes here from the number of ligands that has to be analyzed for the protein of interest. Ligand libraries can hold few millions of members. Calculation of their interaction profiles with the macromolecules is independent between library members. High-throughput computer grids or clusters are the most suitable resources for performing this kind of molecular screening. There are several programs available for virtual screening of macromolecule-ligand (or macromolecule-macromolecule) docking profiles. The programs differ either in the search algorithms (systematic or stochastic searches in the torsion angle space, genetic algorithms, molecular dynamics) or in scoring functions used to estimate the interaction energies. Docking is a very popular computational method according the huge number of papers published during the last few years. This is illustrated on Figure 2 where the methodology differences in docking-related papers are summarized. The total number of papers published (PubMed searches, October 2011) is close to 6,000, covering both, experimental and computational docking approaches. There are around 16 computational methods referenced for docking. The main difference in these programs comes from searching algorithms and scoring functions as mentioned above. We implemented and tested several of them: Dock [2 , 3], AutoDock [4, 5], Glide [6], AutoDock Vina [7], FlexX [8] and HEX [9], although only four docking programs were shown to be used preferentially according the number of publications.

New computational horizons were opened on HW level by introduction of the Graphics Processing Unit (GPU). Recently, selected SW protocols for molecular modeling, computer-aided drug design (CADD) and computer-aided nanomolecular design (CAND) have been updated to benefit from GPU. Many-body interatomic interactions are evaluated by all-atom simulation techniques in MC or MD simulations. These interactions can be described either according to first principles, e.g. Quantum Monte Carlo (QMC), or can be simplified to empirically parametrized interaction potentials used in the majority of MD simulations. The speedup of simulation methods involving hundreds of thousands of atoms has cardinal importance here and can be facilitated by GPU technologies. Indeed, we recognized an important speedup for the GPU version of the above mentioned HEX program for protein-protein docking. More detailed description of the programs benefiting from GPU is summarized in the next paragraph.

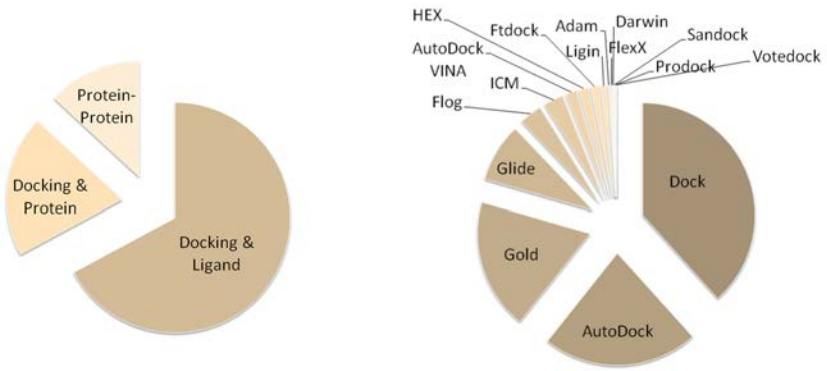

Fig. 2. Results of the PubMed searches for "docking". **Left:** Comparison of all docking studies (both, experimental and computational) with ligand docking into protein's active site resp. docking of a smaller protein into a larger one. **Right:** Comparison of the number of publications using different docking protocols. Four methods were referenced in the majority of articles: Dock [2 , 3], AutoDock [4, 5], Gold [10, 11] and Glide [6]. Only the first two programs (Dock, AutoDock) among the most referenced are non-commercial (academic) ones .

2 Overview of GPU-updated methods for molecular modeling and design

Significant effort was devoted to build a state-of-the-art computational laboratory equipped with the newest GPU-related software releases in order to match the recent trends in computer modeling of biomacromolecules [1] and nanomaterials [12]. Selections of some programs are available for download from several public domains as well as proprietary web pages. Our attention here is restricted to those that have already been updated for GPU computing [1]. NVIDIA published the list of the MD programs with GPU option at
http://www.nvidia.com/object/molecular_dynamics.html.

ACEMD [13] is a commercial molecular dynamics simulation package optimized for GPU environment. The AMBER 11 [14] program suite was very recently also upgraded for GPU.

The NAMD [15] and VMD [16] were among the first programs significantly benefiting from the GPU environment. NAMD is typically used for MD simulations of large molecules within explicit solvent box [1].

The latest version of HOOMD (Highly Optimized Object-oriented Many-particle Dynamics) [17] is available at http://codeblue.umich.edu/hoomd-blue/index.html. HOOMD was originally proposed for polymer simulations and designed to run in

GPU environment. The recent version of LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator (http://lammps.sandia.gov/index.html) has enhanced GPU/CUDA support as well.

Probably the first QC program written directly for GPU CUDA is the TeraChem software [18-21] commercially available at http://www.petachem.com/. From the other QC programs only GAMESS [22, 23] was recently updated for GPU.

3 Modeling amyloid aggregation of insulin

Under certain physiological condition in living cell, some proteins fail to fold into native functional structure individually and aggregate together to form fibril structures, often called amyloids. Accumulation of such fibrils in living cells can initiate severe diseases. Alzheimer´s and Parkinson´s diseases, type II diabetes, dialysis-related amyloidosis or various forms of systemic amyloidosis [24, 25] belong to the amyloid-related diseases. Such systems are being extensively studied both experimentally and through computer simulations [26, 27].

A review of computer modeling studies of nanomaterials in biological environment was recently published by Yarovski et al. [12]. Conformational changes of proteins and nucleic acids induced by nanoparticles may result in amyloid aggregation [12]. The effect of nanoparticles on amyloid aggregation was experimentally studied also at our department. Moreover, small molecules were identified that can efficiently influence the amyloid aggregation. In order to better understand the effect of these molecules we performed "*In Silico*" binding studies.

At first the insulin native structure (Figure 3 top left) and the unfolded dimer (top right) were optimized. Determination of the possible binding sites of these structures was the next step in the computational protocol. This was done using the SiteMap [28] program of Schrodinger LLd. The same molecules that were used in the experimental part of the study [29] were then docked into the binding sites using the AutoDock Vina program [7]. Although this program did not belong to the "preferential" docking choices (see Figure 2), it outperforms the "parent" AutoDock. The reason for smaller usage is simple: the novel AutoDock Vina methodology was published very recently in 2010. In comparison to the more older and popular AutoDock the docking efficiency was significantly improved from 49% (AutoDock) to 78% (Autodock Vina) as was shown on the training set.

The results of the computational modeling of small molecule binding to insulin/native structure versus insulin/amyloid aggregate were very promising. We were able to identify the best binder in global agreement with the experimental binding study. This confirmed that both, the protein model with the calculated binding site as well as the Vina method were chosen properly to model the binding of the small organic molecules, namely phytoalexins.

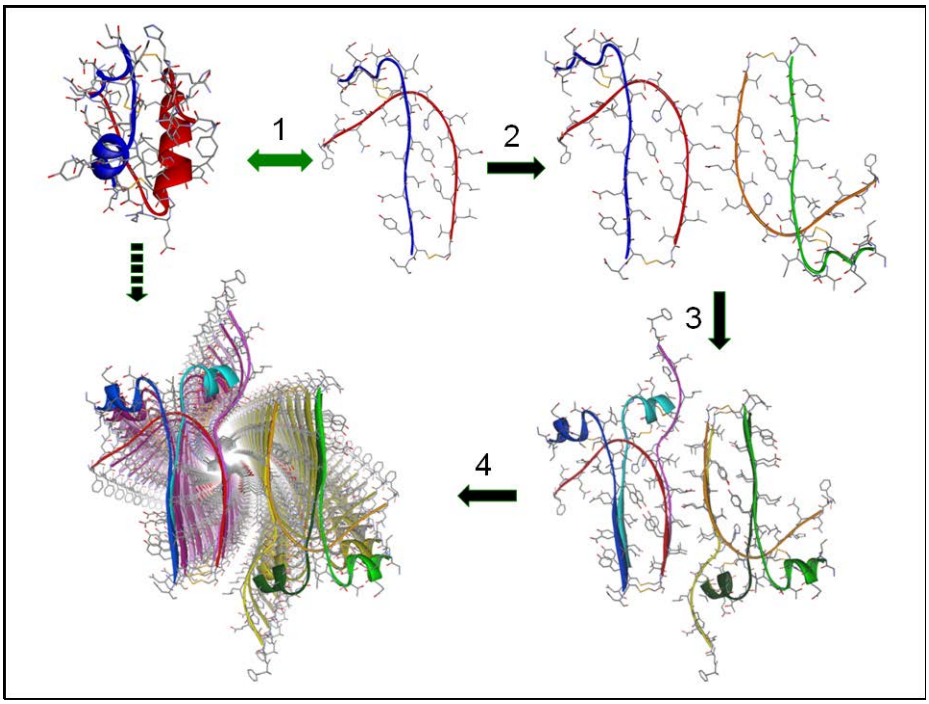

Fig. 3. Possible mechanism of the amyloid aggregation of insulin. **1**. The helical segment of the A (red) and B (blue) chain can unfold into elongated sheet-like monomer. The folding/unfolding can be a reverse procedure. **2**. The monomer can dimerize with "zipper"-like stabilization of the amino acid side chains. **3**. Further stabilization comes from sandwich-like structure. **4**. Equivalent stabilization appears during the fibrillization resulting in extended insulin amyloid aggregate. The coordinates for the starting amyloid fibril structures were downloaded from http://people.mbi.ucla.edu/sawaya/jmol/fibrilmodels.

4 *"In Silico"* arrays in protein-ligand interaction studies

All our preliminary testings of AutoDock Vina indicated that the program can reasonably predict the binding affinities of small molecules. Accordingly, we extended our study towards *"In Silico"* lectin arrays and performed huge number of docking calculations.

We constructed the array as follows: one dimension of the array was relatively small and was composed from 6 lectin molecules, Galectin-1, Galectin-3, Galectin-4, Galectin-7, Galectin-8 and Galectin-9. Galectins belong to the class of galactose binding lectins. The other dimension of the array was significantly larger and was composed from libraries of small molecules. These libraries belong to the ZINC database and we were using the following subsets: Asinex with around 370,000 compounds, Otava (~330,000 compounds) and the database of Natural products (~90,000 compounds). This in total is close to 790,000 compounds. The virtual screening thus needed to run around 4,740,000 computations. Such large high-throughput computa-

tion task was accomplished using the Torque/Maui batch system. In addition, there were 9 geometries stored in each file resulting from successful docking. We used a special (self written) software in order to sort and analyze such a huge dataset and to select the best binders for each galectin under study. The three database sets, Asinex, Otava and the Natural compounds were analyzed separately.

Figure 4 illustrates part of the accomplished results. The best binders from the Natural products database are visualized here. Although the six galectins have similar binding sites, three of them (Galectin-3, Galectin-4 and Galectin-8) bind different small molecules. The other three galectins do not exhibit selectivity in the binding. The same small molecule is the preferred binder for Galectin-1, Galectin-7 and Galectin-9. Further modeling studies are required here to modify the small molecule in order to gain larger binding selectivity.

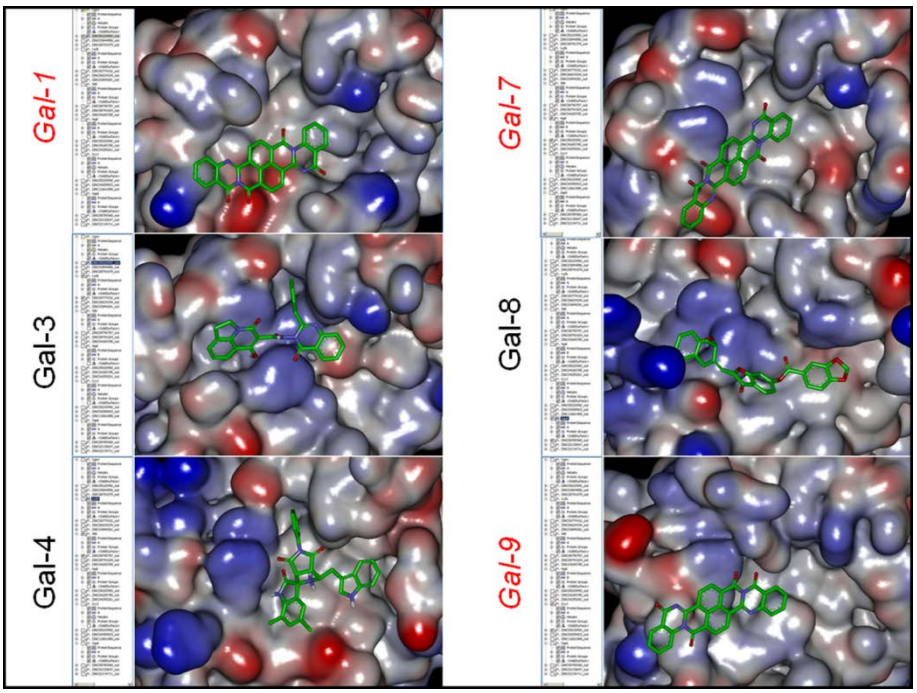

Fig. 4. The best binders to the Galectins under study as resulted from the database of the Natural Compounds. The galectins where the same ligand is the best binder are highlighted red in italic. The van der Waals protein surface of the galectins is colored according electrostatic potential. The ligand atoms are colored as follows: carbon in green, oxygen in red and nitrogen in blue.

5 Conclusions and outlooks

We succeeded to implement, test and use an important set of programs and gain significant speedup resulting from GPGPU. We also succeeded to enhance the efficiency of high-throughput computing by running around 100,000 docking runs/day. In addition to our former efforts to build an efficient international virtual computational laboratory for biomolecular modeling [30], our interest was further expanded overseas. Our recent attention is devoted to establish a joint virtual computation laboratory between Academia Sinica, Taiwan and IEP SAS in Kosice, Slovakia. Within this project, a number of packages developed in the Laboratory of Statistical and Computational Physics (Taiwan), like SMMP [31], ARVO [32], CAVE [33], *etc*. will be upgraded to include the GPU supporting software.

Acknowledgements. The hardware and software resources used for the computations presented in this work were supported by the project 26220120033 provided by the Structural Funds of European Union (SFEU). Additional computational and traveling support comes from the Centre of Excellence of SAS NANOFLUID, the VEGA grant agency (grant 2/0073/10, 2/0079/10), project APVV-0171-10 and from the SAS Slovakia – NSC Taiwan joint research grant.

References

1. Stone, J.E., Hardy, D.J., Ufimtsev, I.S., Schulten, K.: J Mol Graph Model **29**, 116-125 (2010)
2. Good, A.C., Ewing, T.J., Gschwend, D.A., Kuntz, I.D.: J Comput Aided Mol Des **9**, 1-12 (1995)
3. Ewing, T.J., Makino, S., Skillman, A.G., Kuntz, I.D.: J Comput Aided Mol Des **15**, 411-428. (2001)
4. Goodsell, D.S., Morris, G.M., Olson, A.J.: J Mol Recognit **9**, 1-5 (1996)
5. Morris, G.M., Goodsell, D.S., Huey, R., Olson, A.J.: J Comput Aided Mol Des **10**, 293-304 (1996)
6. Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., Mainz, D.T.: J Med Chem **49**, 6177-6196 (2006)
7. Trott, O., Olson, A.J.: J Comput Chem **31**, 455-461 (2010)
8. Hindle, S.A., Rarey, M., Buning, C., Lengaue, T.: J Comput Aided Mol Des **16**, 129-149 (2002)
9. Ritchie, D.W.: Proteins **52**, 98-106 (2003)
10. Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., Taylor, R.D.: Proteins **52**, 609-623 (2003)
11. Nissink, J.W., Murray, C., Hartshorn, M., Verdonk, M.L., Cole, J.C., Taylor, R.: Proteins **49**, 457-471 (2002)
12. Yarovsky, I., Makarucha, A.J., Todorova, N.: Eur Biophys J Biophys Lett **40**, 103-115 (2011)
13. Harvey, M.J., Giupponi, G., De Fabritiis, G.: J Chem Theory Comput **5**, 1632-1639 (2009)
14. Case, D.A., Darden, T.A., T.E. Cheatham, I., Simmerling, C.L., Wang, J., Duke, R.E., R.Luo, Walker, R.C., Zhang, W., Merz, K.M., Roberts, B., Wang, B., Hayik, S., A.

Roitberg, Seabra, G., Kolossváry, I., Wong, K.F., Paesani, F., Vanicek, J., Wu, X., Brozell, S.R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Cui, G., Roe, D.R., Mathews, D.H., Seetin, M.G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., Kollman, P.A.: Amber 11. University of California, San Francisco (2010)
15. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., Schulten, K.: J Comput Chem **26**, 1781-1802 (2005)
16. Humphrey, W., Dalke, A., Schulten, K.: J Mol Graph **14**, 33-38, 27-38 (1996)
17. Anderson, J.A., Lorenz, C.D., Travesset, A.: J Comput Phys **227**, 5342-5359 (2008)
18. Martinez, T.J., Ufimtsev, I.S.: J Chem Theory Comput **4**, 222-231 (2008)
19. Ufimtsev, I.S., Martinez, T.J.: Comp. Sci. Eng. **10**, 26-34 (2008)
20. Ufimtsev, I.S., Martinez, T.J.: J Chem Theory Comput **5**, 1004-1015 (2009)
21. Martinez, T.J., Ufimtsev, I.S.: J Chem Theory Comput **5**, 2619-2628 (2009)
22. Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S.: J Comput Chem **14**, 1347-1363 (1993)
23. Gordon, M., Schmidt, M.: Theory and Applications of Computational Chemistry: the first forty years, CE Dykstra, G. Frenking, KS Kim, GE Scuseria (editors) 1167–1189
24. Westermark, P., Wernstedt, C., Wilander, E., Sletten, K.: Biochem Biophys Res Commun **140**, 827-831 (1986)
25. Soto, C.: Nat Rev Neurosci **4**, 49-60 (2003)
26. Li, M.S., Co, N.T., Reddy, G., Hu, C.K., Straub, J.E., Thirumalai, D.: Phys Rev Lett **105**, (2010)
27. Berhanu, W.M., Masunov, A.E.: J Mol Model (2011)
28. Halgren, T.: Chem Biol Drug Des **69**, 146-148 (2007)
29. Siposova, K., Antosova, A., Kutschy, P., Daxnerova, Z., Fedunova, D., Bagelova, J., Kozar, T., Gazova, Z.: submitted (2011)
30. Komaromi, I., Toth, L., Kozar, T.: Computational "virtual laboratory" tools for biomolecular and drug design. In: Hluchý, L., Sebestyénová, J., Kurdel, P., Dobrucký, M. (eds.): 4th International Workshop on Grid Computing for Complex Problems. Institute of Informatics, Slovak Academy of Sciences, Bratislava 86-93 (2008)
31. Eisenmenger, F., Hansmann, U.H.E., Hayryan, S., Hu, C.K.: Comput Phys Commun **138**, 192-212 (2001)
32. Busa, J., Dzurina, J., Hayryan, E., Hayryan, S., Hu, C.K., Plavka, J., Pokorny, I., Skrivanek, J., Wu, M.C.: Comput Phys Commun **165**, 59-96 (2005)
33. Hu, C.K., Busa, J., Hayryan, S., Skrivanek, J., Wu, M.C.: Comput Phys Commun **181**, 2116-2125 (2010)