Docking and Virtual Screening Using Distributed Grid Technology

Alfonso T. García-Sosa,* Sulev Sild, and Uko Maran

Institute of Chemistry, University of Tartu, Jakobi 2, Tartu 51014, Estonia
*E-mail: alfonsog@ut.ee

Keywords: H5N1 Avian Influenza, HIV-1, Reverse Transcriptase, Grid computing, Quantitative structure-activity relationships, Toxicity, Bioinformatics, Drug design

Presented at the EuroQSAR Conference

Received: December 10, 2008; Accepted: May 3, 2009

DOI: 10.1002/qsar.200810174

Abstract

Distributed grid technologies are gradually realizing their potential to provide innovative infrastructures for complex scientific and industrial applications in the field of computational chemistry and related application areas. The current paper gives examples of distributed solutions for docking and virtual screening applications in moving the computational paradigms towards collaborative research and grid computing environments. The Chemomentum collaborative computing environment including both hardware and software infrastructure is described. Examples of applications are given i) for docking and virtual screening on multiprotein and multilibrary cases for H5N1 avian influenza and HIV-1 viruses; and ii) QSAR model building related to HIV-1 protease activity and aquatic toxicity.

1 Introduction

Grid and distributed computing have been increasingly used for over a decade, using large computer resources to solve questions of medical or biological significance, such as the well-known Screensaver Lifesaver [1], the Folding@ home project [2] and the Aids@Home [3] to mention a few. Massively parallel supercomputers are very effective at solving complex scientific problems [4–7]. One of the main advantages of grid computing is a readily available software infrastructure that can harness computer resources at times of low demand or for collaborative searches where extensive CPU/hours are needed. In addition, distributed chemical applications such as web-based services for chemoinformatics [8, 9] and Java remote method invocation for the calculation of molecular descriptors, have been described [10]. Cloud computing is another new development where computer resources are mingled and made accessible over the internet [11].

Grid computing is a viable tool for addressing many chemistry related applications. The most common type of grid service is to provide a distributed access for running various software packages. In this case, the grid is used as a user-friendly front end to submit computationally intensive jobs for available computer resources. A number of examples have been previously summarized elsewhere [12, 13]. More recent examples include the calculation of protein-ligand binding affinities [14, 15], virtual screening for plasmepsin inhibitors on the EGEE grid infrastructure in order to prevent malaria [16], and molecular dynamics simulations with CHARMM software on the Open Science Grid [17].

Another common application type in grids is distributed access to chemistry related databases. For example, the OpenMolGRID project has developed a data warehouse with relevant data for QSAR modeling [12] and the U.K. National Crystallography Service has developed a grid based e-science infrastructure for small molecule crystallography services [18]. The Common Instrument Middleware Architecture provides grid services to control scientific instruments and to access their data [19].

An interesting use of grid technology is carrying out scientific workflows that combine distributed application resources and data sources for solving complex tasks. This is practical because many scientific workloads involve the execution of multiple applications in a predefined sequence. Relevant examples include the OpenMolGRID [12], SOMA [20] and GEMSTONE [21] systems. A new example of a workflow centric system oriented to molecular design and modeling is Chemomentum [22].

2 Computational Environment

Chemomentum [23] is an open, collaborative environment for grid computing which utilizes the open source UNICORE grid middleware [24, 25]. Its predecessors were the OpenMolGRID system [12] and BioGRID project [26]. It is a workflow-centric system that allows the precise specification and reproduction of in silico screening protocols (e.g. molecular docking or QSAR prediction), and includes many applications for computational modeling tasks. At the moment of writing, it covers more than twenty: data filtering and preprocessing, classification of chemicals, generation of 3D atomic coordinates, predictive model building (linear and nonlinear, including prediction), molecular descriptor calculation, quantum chemistry, conformational space analysis, genetic sequence analysis, molecular dynamics, and molecular docking. It puts efforts in semantics, metadata and knowledge management in distributed and heterogeneous environments in order to build up a documented data space as a system for data storage, retrieval and exchange [27]. It also can create a virtual organization that joins the expertise and resources of different institutions and groups, allowing strict licensing schemas and data access where computational resources (both hardware and software) are appropriately pooled and made available. The system installations can be organization-specific in small or large scale.

Whereas many research activities in chemistry and drug design involve separate pieces of software and file formats, they can be integrated and ran together in a single workflow in Chemomentum. For example, a QSAR model prediction for a compound dataset where 3D geometry conversion from 2D coordinates is done, followed by quantum chemical and molecular descriptors calculations, and finally building the QSAR model, can all be setup in a single automated workflow (Fig. 1). In addition, all the programs may reside on different machines which reduce the need for installation, as well as platform problems arising from different machine architectures, etc. UNICORE software can run in multiple environments and there are several clients available, such as command line and web clients, as well as graphical user interfaces based on the Eclipse platform.

Chemomentum does require a certain involvement of the user site in the installation of the UNICORE and up-to-date Java software, as well as downloading and having security certificates assigned and verified by the administrators. It may also entail the reciprocal setting aside of a small, ring-fenced portion of one's own computing resources for other partners in the consortium to use, as well as reporting bugs and wish lists to the developers for further development of this scientific software.

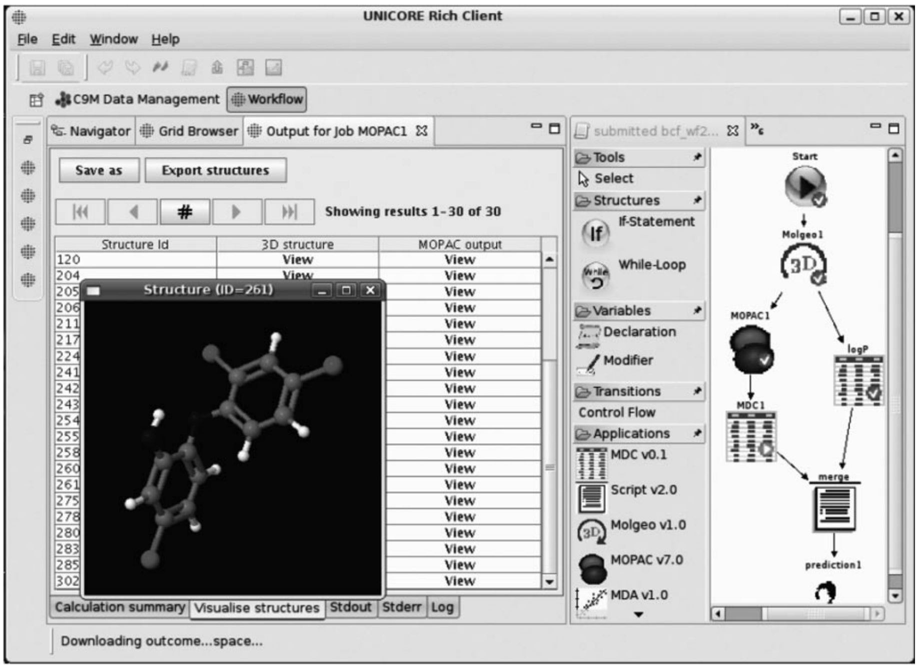
3 Applications and Discussion

3.1 Virtual Screening

Viruses are of growing importance in healthcare. They commonly have fast mutating components such as enzymes. Given several structures of recently identified protein targets, docking and virtual screening were carried out on multiprotein, multilibrary cases for HIV-1 and H5N1 influenza viruses. The results were confirmed using agreements between different docking programs (Glide [28] and Autodock [29]) and different methods (consensus screening, ligand efficiency). For both screening cases, nearly 70,000 structures were used from filtering compounds from the NCI [30], DrugBank [31] and ZINC 7 [32] databases, and our consensus approach required that a ligand be scored highly by each docking program separately in order for a compound to receive a high rank. In addition, ligand efficiency measures were employed such as dividing the free energy of interaction by the number of heavy atoms, or by the molecular weight of the ligand, in order to characterize compounds.

3.2 H5N1 Avian Influenza

H5N1 avian influenza is a recently evolved, deadly disease, which has credible potential to transform into a human pandemic [33]. Neuraminidase (EC 3.2.1.18, also called sialidase) is a hydrolase enzyme that is crucial to the entry and release of the virus from a host cell membrane. Screening was conducted against the crystal structures of a wild-type neuraminidase (2HU0.pdb) and of an oseltamivir-resistant mutant found in EU populations (3CKZ) [34] as well as from molecular dynamics simulations of the wild-type protein [35]. The resulting bound complexes showed ligands engaged in similar interactions with the protein as those of known drugs oseltamivir and zanamivir (Fig. 2). Among the top binders there are molecules with a strong resemblance to saccharides, as well as novel chemical classes. This is encouraging since the natural substrates of neuraminidase are glycolipids and glycoproteins. These molecules have a complex stereochemistry and flexibility that can be exploited for drug design. Amplifying the stereocenters can provide new compound diversity, and the flexibility may allow some of these compounds to adapt to viral residue mutations or conformational changes in the binding sites. Beta-lactams were also among high-ranked compounds against the wild-type protein. However, they were no longer present as the top-binders against the drug-resistant mutant protein. Neither was oseltamivir, in agreement with experiment [34] as was also the fact that zanamivir is a top-binder. Table 1 shows the structure of 2 promising candidates, as well as 2 known inhibitors, together with their calculated consensus interaction energies.



UNICORE Rich Client

File Edit Window Help

C9M Data Management | Workflow

Navigator | Grid Browser | Output for Job MOPAC1

Save as | Export structures

Showing results 1-30 of 30

Structure Id	3D structure	MOPAC output
120	View	View
204	View	View
205		View
206		View
211		View
217		View
224		View
241		View
242		View
243		View
254		View
255		View
258		View
260		View
261		View
275		View
278		View
280		View
283		View
285		View
302		View

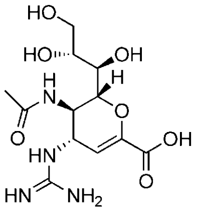
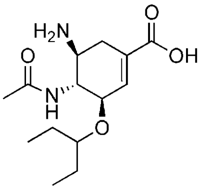
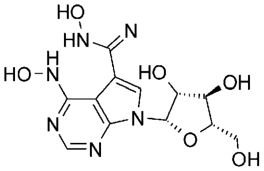
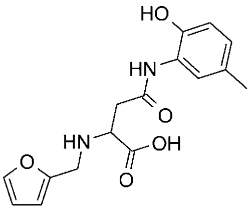
Structure (ID=261)

Calculation summary | Visualise structures | Stdout | Stderr | Log

Downloading outcome...space...

Figure 1. Multistep workflow image in Chemomentum.

Table 1. Chemical structure, consensus score interaction energies (I.E., in kcal/mol), and ligand efficiencies (I.E./number of heavy atoms, and I.E./molecular weight) of inhibitors and candidates against wild-type and drug-resistant H5N1 neuraminidase.

Name or ID	Structure	Interaction with wild-type protein	Interaction with drug-resistant protein
Zanamivir		– 8.00, – 0.348, – 0.024	– 9.42, – 0.449, – 0.028
Oseltamivir		– 7.26, – 0.363, – 0.026	– 7.58, – 0.380, – 0.027
NSC154829		– 8.41, – 0.381, – 0.027	–
ZINC05258197		–	– 9.57, – 0.456, – 0.030

Several ligand efficiencies[36] were calculated for these molecules which may explain their activity, since they have a strong interaction energy per heavy atom. Some of the highly ranked compounds formed complexes with both forms of neuraminidase [37] in 3 binding sites simultaneously, including conserving a carboxylate from oseltamivir in the original sialic acid binding site and its interactions with the protein (Fig. 2).

Optionally, the flexible parts of the molecules can be rigidified with the introduction of a double bond, for example, in order to increase affinity by reduction of entropy loss on binding. The moderately polar nature of the compounds shows that they could act in the extra-cellular environment where neuraminidase is active.

3.3 HIV-1 Reverse Transcriptase

HIV-1 was isolated in the beginning of the 1980's, has claimed 25 million lives, and affects 30 million people today [38]. Reverse transcriptase (EC 2.7.7.49) is one of the main target enzymes for HIV-1 since it is only produced by the virus and has no analogous protein in the human organism, and is used by the virus when transcribing its single-stranded RNA into double-stranded DNA.

Several reverse transcriptase structures were used, both wild-type (1S9E.pdb) and drug-resistant (2IC3.pdb). Among the top binders are compounds that have a high resemblance to efavirenz (Fig. 3), which is used therapeutically for the treatment and suppression of HIV-1 infection in combination with other nonnucleoside reverse-transcriptase inhibitors (NNRTI) drugs, and as the first line treatment in preference to protease inhibitors. These structures are also from different chemical classes (see ZINC5932839 and ZINC00018716 in Table 2).

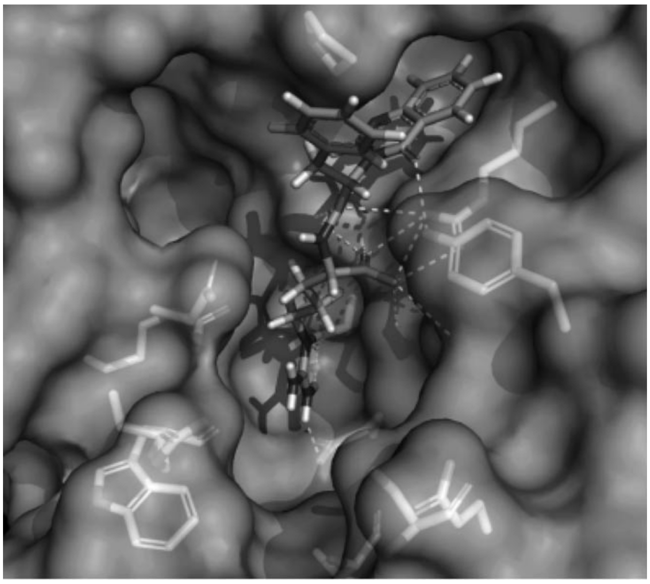


Figure 2. H5N1 avian influenza neuraminidase in complex with a docked predicted inhibitor spanning multiple binding sites. Original oseltamivir and sialic acid binding site bottom middle, hydrogen bonds are indicated by dashes.

Table 2 shows the structures of 2 promising candidates against both wild-type and drug-resistant forms of the HIV-1 enzyme, as well as a known nonnucleoside inhibitor against reverse transcriptase (efavirenz), together with their calculated consensus interaction energies. It is interesting to note the similarity of efavirenz with ZINC00880460 (see Table 2). Both compounds share a similar bicyclic system, with substitutions in close positions on the phenyl ring for both molecules, in addition to a similar distribution of amide and other polar groups on the first heteroatomic ring that participate in hydrogen bonding with the protein, as well as large hydrophobic groups substituting in close positions on this ring.

3.3.1 Discussion on Chemomentum

We ran Autodock through the Unicore command line client, connecting to different international sites included in the service registry of the Chemomentum system. Running through the command line client requires only the setup of input and output files for batch processing. In addition, job description files are required with the instructions on which application service to use, what arguments to pass to it, and the input/output locations. The jobs were generally run fast, (average time per docking run per ligand per protein was around 18 min.) in parallel over the distributed system and with a positive outcome, although a few jobs (out of thousands) failed and had to be resubmitted. Chemomentum is particularly well suited for open or academic-licensed software (such as Autodock) which can benefit from a large number of processors, since it can then be deployed over many systems at once in which all sites have the authorization to run the program. If a smaller calculation project is needed, or if the license model does not allow the described setup, a simpler, user-centric UNICORE installation may also be built, or a smaller, local queuing system may be preferred instead of Chemomentum. For example, the Glide docking jobs described were run on our local system.

3.4 Quantitative Structure-Activity Relationships

QSAR is a well-established ligand based technique to predict properties of compounds such as activity, solubility and toxicity, among others. A strong benefit of using QSAR in the workflow centric distributed environment is the ability to precisely set and document the protocol which eases the reproducibility of the in silico model, due to the support of metadata and the archiving of software settings that allows to conveniently reuse the workflows.

QSAR models were built to predict the activity of compounds against HIV-1 aspartic protease[39] (EC 3.4.23), and to predict toxicity in aquatic species. Multiple linear regression (MLR) and artificial neural networks (ANN)

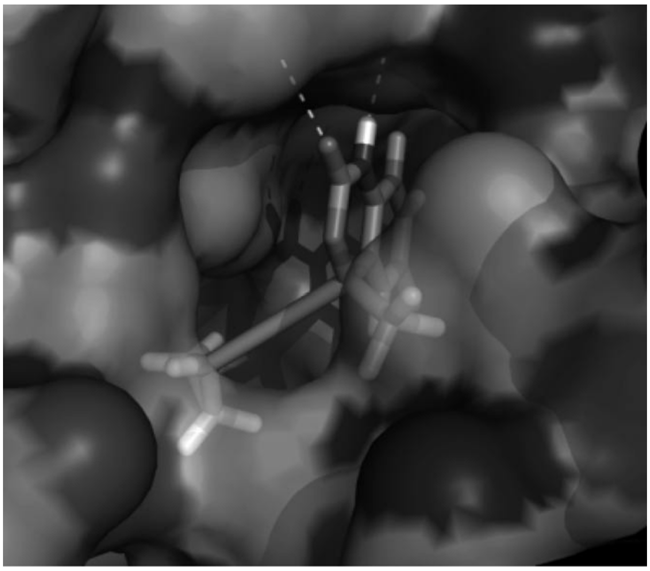


Figure 3. Complex of HIV-1 reverse transcriptase with docked inhibitor efavirenz, hydrogen bonds are indicated by dashes.

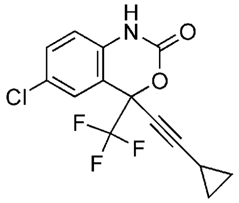
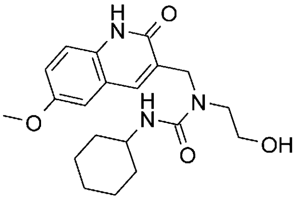
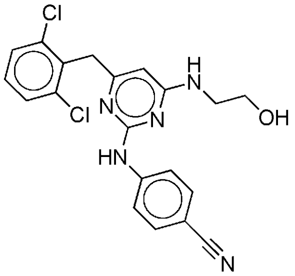
were studied comparatively both for the unicellular organism *T. pyriformis* [40], as well as for the fish *Pimephales promelas* [41]. In both cases ANN showed higher prediction power than the MLR model.

3.4.1 Discussion on Chemomentum

The experience from carrying out QSAR modeling with the Chemomentum system in grid environment has been positive. Usually the main motivation for grid computing is to speed up computation. The QSAR model development for small data sets usually does not demand significant computing power. However, the analysis of large data sets, the calculation of quantum chemical descriptors, and the training of artificial neural networks demands large amounts of computing resources. Currently, the Chemomentum system can automatically split up data sets and distribute them across available resources for speeding up the molecular descriptor calculation process. The speedup of model development is not possible without the redesigning of model development software to exploit distributed parallel processing. Therefore, the speedup of the overall modeling process greatly depends on the descriptor calculation methodology used.

Another important feature in the Chemomentum system is the automation of the QSAR model development

Table 2. Chemical structure, consensus score interaction energies (I.E., in kcal/mol), and ligand efficiencies (I.E./number of heavy atoms, and I.E./molecular weight) of the known inhibitor efavirenz and candidate inhibitors against both wild-type and drug-resistant HIV-1 reverse transcriptase.

Name or ID	Structure	Interaction with wild-type protein	Interaction with drug-resistant protein
Efavirenz		– 12.87, – 0.613, – 0.040	– 11.35, – 0.540, – 0.036
ZINC0880460		– 11.89, – 0.440, – 0.032	– 12.08, – 0.447, – 0.032
ZINC5932839		– 13.30, – 0.475, – 0.032	– 12.12, – 0.433, – 0.029

process. The automation reduces the amount of manual work a user needs to perform. This saves time and reduces the chance of human errors. Of course, the full automation is not always possible and some manual intervention is still required. For example, sometimes data sets contain a few compounds that require multiple trials to get the input geometry right for the quantum chemical calculations. Workflows are very useful for testing and comparing different modeling approaches. Workflow tools have high relevance outside grid computing, for example, the SciTegic Pipeline Pilot is used for QSAR modeling [42].

Finally, the Chemomentum system encourages the collaborative work between scientists. Data sets can be securely stored in the data management system so that multiple users can share the same data. The final and intermediate results produced from workflows are stored in the data management system and are ready for sharing with colleagues. The collaborative aspect in grid computing is very useful and is definitely worthwhile for further developments.

4 Conclusions

Grid-based applications are increasingly featured in scientific studies. Workflows can be set-up and reused easily, with the resulting benefits for reproducibility and productivity. Less time is spent installing different programs on separate architectures and environments, spread over several locations. Open and collaborative environments such as the one briefly described here, can be useful for pooling computational resources, chemistry tools and scientific knowledge and expertise. The advantage of such collaborations is valuable in preventing mutating and fast spreading viral diseases. It allows effective processing of compound libraries for docking and screening against a number of proteins, and building and feeding into QSAR models for predicting a variety of molecular properties, such as ADME profile, solubility, toxic side effects, etc. Ultimately, it may aid in the fast development of vaccines, therapeutics, diagnostics, and/or to predict whether the strain of the virus can be resistant to known drugs, or to aid health policy related decisions.

5 Acknowledgement

We thank the EU FP6 Chemomentum Project (grant IST-033437), the Estonian Science Foundation (grants JD80, 7099, and 7153) and the Estonian Ministry for Education and Research (grant SF0140031Bs09) for funding, Mr. Indrek Tulp and Mr. Kalev Takkis for contributions to the filtering of compounds, and University of Zurich, Istituto Mario Negri (Milan, Italy), Research Centre Jülich, and Technical University Dresden (both Germany) for additional computing resources.

6 References

[1] W. G. Richards, *Nature Rev. Drug Discovery* **2002**, *1*, 551–555.
[2] M. R. Shirts, V. Pande, *Science* **2000**, *290*, 1903–1904.
[3] http://fightaidsathome.scripps.edu (accessed 14 Apr. 2009).
[4] M. L. Klein, W. Shinoda, *Science* **2008**, *321*, 798–800.
[5] A. N. Lupas, *J. Struct. Biol.* **2008**, *163*, 254–257.
[6] K. Y. Sanbonmatsu, C.-S. Tung, *J. Struct. Biol.* **2007**, *157*, 470–480.
[7] A. T. García-Sosa, M. Castro, *Int. J. Quantum Chem.* **2000**, *80*, 307–319.
[8] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko, *J. Comput.-Aided. Mol. Des.* **2005**, *19*, 453–463.
[9] X. Dong, K. E. Gilbert, R. Guha, R. Heiland, J. Kim, M. E. Pierce, G. C. Fox, D. J. Wild, *J. Chem. Inf. Model.* **2007**, *47*, 1303–1307.
[10] M. Karthikeyan, S. Krishnan, A. K. Pandey, A. Bender, A. Tropsha, *J. Chem. Inf. Model.* **2008**, *48*, 691–703.
[11] http://www.acm.org/ubiquity/volume_9/v9i31_delic.html (accessed 2 Dec. 2008)
[12] S. Sild, U. Maran, A. Lomaka, M. Karelson, *J. Chem. Inf. Model.* **2006**, *46,* 953–959.
[13] P. Bała, K. Baldridge, E. Benfenati, M. Casalegno, U. Maran, Ł. Mirosław, Vitaliy Ostropytskyy, K. Rasch, S. Sild, R. Schöne, B. Schuller, N. Williams in: *Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare* (Ed: M. Cannataro) **2008**, ch. XXX, pp. 1–29.
[14] S. P. Brown, S. W. Muchmore, *J. Chem. Inf. Model.* **2006**, *46*, 999–1005.
[15] P. W. Fowler, S. Geroult, S. Jha, G. Waksman, P. V. Coveney, *J. Chem. Theory Comput.* **2007**, *3*, 1193–1202.
[16] V. Kasam, M. Zimmermann, A. Maaß, H. Schwichtenberg, A. Wolf, N. Jacq, V. Breton, M. Hofmann-Apitius, *J. Chem. Inf. Model.* **2007**, *47*, 1818–1828.
[17] A. Damjanovic, B. T. Miller, T. J. Wenaus, P. Maksimovic, B. Garcia-Moreno E., B. R. Brooks, *J. Chem. Inf. Model.* **2008**, *48*, 2021–2029.
[18] S. J. Coles, J. G. Frey, M. B. Hursthouse, M. E. Light, A. J. Milsted, L. A. Carr, D. DeRoure, C. J. Gutteridge, H. R. Mills, K. E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke, M. Day, *J. Chem. Inf. Model.* **2006**, *46*, 1006–1016.
[19] R. Bramley, K. Chiu, T. Devadithya, N. Gupta, C. Hart, J. C. Huffman, K. Huffman, Y. Ma, D. F. McMullen, *J. Chem. Inf. Model.* **2006**, *46*,1017–1025.
[20] P. T. Lehtovuori, T. H. Nyrönen, *J. Chem. Inf. Model.* **2006**, *46*, 620–625.
[21] K. Baldridge, K. Bhatia, B. Stearn, J. P. Greenberg, S. Mock, S. Krishnan, W. Sudholt, A. Bowen, C. Amoreira, Y. Potier, in *Grid Computing in Life Sciences* (Eds: T. T. Wee, P. Arzberger, A. Konagaya), *LSGRID 2005*, World Scientific, Singapore **2006**, pp. 155–175.
[22] *Chemomentum – Grid Services Based Environment to Enable Innovative Research*, http://www.chemomentum.org (accessed 3 Dec. 2008)
[23] B. Schuller, B. Demuth, H. Mix, K. Rasch, M. Romberg, S. Sild, U. Maran, P. Bała, E. del Grosso, M. Casalegno, N. Piclin, M. Pintore, W. Sudholt, K. K. Baldridge, in (Eds: L. Bougé, et al.) *Euro-Par 2007 Workshops: Parallel Processing*, Springer, Berlin **2008**, LNCS 4854, pp. 82–93.

[24] *Distributed Systems and Grid Computing*, Jülich Supercomputing Centre, Institute for Advanced Simulation, Research Centre Jülich, Germany, UNICORE-Distributed computing and data resources, http://www.unicore.eu (accessed 15 Jan. 2008).

[25] J. Almond, D. Snelling, *Future Gener. Comput. Sys.* **1999**, *613*, 439 – 548.

[26] J. Pytlinski, L. Skorwider, P. Bala, M. Nazaruk, K. Wawruch, in *Euro-Par 2002: Proc. 8th Int. Euro-Par Conf. on Parallel Processing* (Eds: B. Monien, R. Feldman), Springer, Berlin **2002**, LNCS 2400, pp. 881 – 884.

[27] K. Rasch, R. Schöne, V. Ostropytskyy, H. Mix, M. Romberg, in (Eds: A. Streit et al.) *Euro-Par 2008 workshops: Parallel Processing*, Springer, Berlin **2009**, LNCS 5415, pp. 84 – 93.

[28] *Glide*, version 4.5, Schrödinger, LLC, New York **2007**.

[29] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem*. **1998**, *19*, 1639 – 1662.

[30] National Cancer Institute/National Institutes of Health USA, http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html (accessed 15 Jan 2008)

[31] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, *34*, D668 – D672.

[32] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.

[33] *Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO, 18 Mar 2008*, World Health Organization **2008**.

[34] P. J. Collins, L. F. Haire, Y. P. Lin, J. Liu, R. J. Russell, P. A. Walker, J. J. Skehel, S. R. Martin, A. J. Hay, S. J. Gamblin, *Nature* **2008**, *453*, 1258 – 1261.

[35] R. E. Amaro, D. D.L. Minh, L. S. Cheng, W. M. Lindstrom, A. J. Olson, J.-H. Lin, J. A. McCammon, *J. Am. Chem. Soc.* **2007**, *129*, 7764 – 7765.

[36] C. Hetényi, U. Maran, A. T. García-Sosa, M. Karelson, *Bioinformatics* **2007**, *23*, 2678 – 2685.

[37] A. T. García-Sosa, S. Sild, U. Maran, *J. Chem. Inf. Model.* **2008**, *48*, 2074 – 2080.

[38] Joint United Nations Programme on HIV/AIDS, Report on the global AIDS epidemic, 2006.

[39] U. Maran, S. Sild, I. Kahn, K. Takkis, *Int. J. Grid Computing Theory, Meth. Appl.* **2007**, *23*, 76 – 83.

[40] I. Kahn, S. Sild, U. Maran, *J. Chem. Inf. Model*. **2007**, *47*, 2271 – 2279

[41] U. Maran, S. Sild, P. Mazzatorta, M. Casalegno, E. Benfenati, M. Romberg, in: *Grid Computing in Computational Biology* (Eds: W. Dubitzky et al.), GCCB 2006, Springer, Berlin **2007**, LNBI 4360, pp. 60 – 74.

[42] S*citegic Pipeline Pilot*, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123 – 1365, U.S.A. Available from SciTegic Inc. at http://accelrys.com/products/scitegic/ (accessed 3 Dec 2008).