

Bridging the computer and life sciences: the case of VI-SEEM

Anastas Mishev

Faculty of Computer Science and Engineering, University Ss Cyril and Methodius, Rudjer Boskovic 16, 1000 Skopje, Macedonia

Introduction

Currently, the world of computing is overwhelmed with huge numbers depicting the extraordinary performance of the current computers. This is especially true for the biggest supercomputers in the world, showing off their exa-FLOPS, aggregating peta-bytes of data, measuring and comparing each other on several top charts (Filiposka et al., 2016). But is this their real power? If we compare the best of the best from only 23 years ago (at the time the first top500 (Top500) list was published), with the today's list leader, there is more than 400000 times increase in the peak performance capabilities. One can have all the exas and petas in the world, but the true power of today's supercomputers is in their applicability to solve real life problems. VI-SEEM (VI-SEEM) is one of the examples of trying to put the supercomputers to work on practical problems. It tries to facilitate regional interdisciplinary collaboration, focusing on the scientific communities of Life Sciences, Climatology and Digital Cultural Heritage. Through unification of the existing e-infrastructure into an integrated platform, it strives to better utilize synergies, for an improved service provision within a unified Virtual Research Environment to be provided to scientific communities of high impact in the combined South East Europe and Eastern Mediterranean region.

The project

Building on the success of the previous regional projects and initiatives that helped bridge the digital divide by ensuring access to regional e-Infrastructures, VI-SEEM includes partners from 16 countries in the SEE and EM region, both from the resource providers' and potential users' communities. Bridging the two worlds would bring new value and improve research productivity and

competitiveness on the pan-European level.

The general project objective is to provide integrated e-Infrastructure platform for regional cross-border Scientific Communities in Climatology, Life Sciences, and Cultural Heritage for the SEEM region that will be user-friendly and accessible to the fore mentioned communities. This goal will be achieved by linking compute, data, and visualization resources, as well as services, models, software and tools. This Virtual Research Environment - VRE will provide the scientists and researchers with the support in full lifecycle of collaborative research: accessing and sharing relevant research data, using it with provided codes and tools to carry out new experiments and simulations on large-scale e-Infrastructures, and producing new knowledge and data - which can be stored and shared in the same VRE. Through training, user support, application development and porting, the researchers will be able to truly utilize the power of the regional e-infrastructure, to try to solve realistic problems, including computer aided drug delivery, modelling of biomolecules, introduction of novel methodologies into drug development, regional genotype databases development are only a few of the possible applications. Our expectations are that this and similar projects will actually bridge the gap between the computing power and its real applications, for a healthier world and better living.

VI-SEEM for life sciences

Advances in computational infrastructure during the last decade have facilitated the development of biological data analysis for big data and computational biology as key research methodologies in both academia and industry. The use of computers in biology has enabled our better understanding of mechanistic aspects in health and disease and has accelerated the development

of novel therapeutics. In this project, the Life Sciences research community is chosen because of its central role in achieving a higher quality of life in the SEEM region. The aim of the VRE is to create and provide the necessary services over a capable infrastructure to facilitate research for understanding of disease mechanisms and appropriate mitigation methodologies in the SEE and EM populations. Project participants and related institutes will assist in data collection and analysis, run and optimizing computational codes and using the research results to understand the molecular basis of diseases associated with SEE and EM areas with projections to develop personalized therapies.

The Life Sciences research community in the SEEM region could benefit greatly from the e-infrastructures at hand. Large amounts of data need to be stored and be made available to researchers for processing in the compute centres of the region. Therefore, apart from storage resources, fast and reliable networking infrastructure is important for moving large datasets from data archives to the computing centres and also moving simulation results to the researchers' facilities for further post processing and acquisition of results. In terms of compute infrastructure, the models and services to be used by the research groups require capacity and capability computing as well as the provision of computing resources for the installation of user facing services. For example, codes such as NAMD and NWCHEM scale up to hundreds or thousands of cores and can benefit from scalable HPC clusters or supercomputers such as the IBMs BlueGene. Molecular dynamics applications are also known to perform well on GPU systems, while also are being ported to new Intel's Phi accelerator platform. On the other hand, parametric codes for human genome sequence analysis can benefit greatly from the Grid or Cloud IaaS computing model. Finally, user-facing services can be also installed in the IaaS infrastructure that will be available in the project. It is evident that the Life Sciences Scientific Community requires a variety of infrastructure resources all of which are going to be available in the VI-SEEM VRE.

Life science use cases

Some most important and most representative examples of using the regional e-Infrastructure for the needs of the Life Science VRE include:

- Modelling and Molecular Dynamics (MD) study of proteins, membrane proteins and biological model membranes. These three biomolecular entities are responsible for signal transduction and are important drug targets. Therefore, in order to design more efficient drugs and drug delivery systems, a better understanding of the physicochemical interactions that govern biomembrane and protein interfaces is needed.
- Computational simulation of DNA and RNA to enable studying the influence of thermodynamic properties of the DNA/DNA and RNA/DNA duplexes on the

transcription and processing of RNA. Computational modelling of the structure, thermodynamics and kinetics of RNA, involved in cancer cell growth

- Computer-aided drug design. By using computational methods and the 3D structural information of the protein target, we are now able to investigate the detailed underlying molecular and atomic interactions involved in ligand: protein interactions and thus interpret experimental results in detail.
- Image processing for biological applications includes experiments by spinning disk confocal microscopy of living cell, which generate images of dozens of GBs per experiment. The generated images require extensive image processing, such as registration, deconvolution, volume rendering, surface rendering, object detection, measurement of shape, size, and intensity of cell objects and automatic object movement tracking of the living cell in 3 dimensions and time.
- Analysis of Next Generation DNA sequencing data to identify disease mechanism pathways and provide patients with timely diagnosis, assessment of risk for developing the disease, targeted and efficient therapy, and give support for possible future reproduction planning.
- Synchrotron data analysis: SESAME is a 3rd generation synchrotron light source that produces very intense pulses of light/X-rays, with wave lengths and intensities that allow detailed studies of objects ranging in size from human cells, through viruses down to atoms, with a precision that is not possible by other means.

The list above is does not limit the possible usage of the resources, only provides some current and ongoing efforts in using the computational, networking and storage infrastructure to aid the Life Science research communities.

The research into computer aided drug design will be given a strong focus during the project, both from the LS researchers, but also from the infrastructure support point of view. Through advances in the drug delivery modelling, novel and hybrid methodologies (Markova et al., 2015) such as molecular dynamics, statistical physics, Monte Carlo etc. will be compared to the traditional methodologies, enabling better understanding of the processes at a very small scale. Through computer aided molecular design (Ng et al., 2015), the simulation results are expected to significantly reduce the clinical trials in anticancer drug research (Kim et al., 2013).

Conclusion

Enabling access to e-Infrastructure through intuitive and user friendly interfaces could bring great benefit to the research communities in the SEE and EM regions. Through virtual collaborative environment, these communities can achieve research excellence on the pan-European and global level. From the point of view of the

e-Infrastructures, strong justification of the investments can be accomplished, demonstrated through real life results. Bridging possibilities of the high end computing infrastructures and the life science scientific communities will produce deeper knowledge of the human biology, better disease understanding, shorter development time of new and targeted drugs with less clinical trials. As in many other cases, the addition of these two will bring much more to the humanity than their simple sum.

Acknowledgment

This work was supported in part by the European Union's Horizon 2020 research and innovation programme, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

References

- Filiposka, S., Mishev, A., Juiz, C., 2016. Current prospects towards energy-efficient top HPC systems. *Computer Science and Information Systems* 13(1), 151–171.
- Top500. Available at: <http://top500.org/>, visited March 2016.
- VI-SEEM. Available at: <http://vi-seem.eu/>, visited March 2016.
- Markova, N., Pejov, Lj., Enchev, V., 2015. A Hybrid statistical mechanics-quantum chemical model for proton transfer in 5-azauracil and 6-azauracil in water solution. *Int. J. Quantum Chem.* 115, 477–485.
- Ng, L.Y., Chong, F.K., Chemmangattuvalappil, N.G., 2015. Challenges and opportunities in computer-aided molecular design. *Computers & Chemical Engineering* 81, 115-129.
- Kim, M., Gillies, R.J., Rejniak, K.A., 2013. Current advances in mathematical modelling of anti-cancer drug penetration into tumour tissues. *Frontiers in Oncology* 3, article 278, 1-10.