3-D electronic density maps for enzyme reaction intermediates: big data and high performance computing services in a virtualized environment

Anton Brekhov, Vladimir Mironov
Department of Chemistry, Lomonosov Moscow State University, Leninskie Gory 1/3, Moscow, 119991, Russian Federation

Alexander Moskovsky
RSC Group, Kutuzovsky prospect 36/23, Moscow, 121170, Russian Federation

Dmitry Podgainiy, Oksana Streitsova, Dmitry Delyakov
Laboratory of Information Technologies, Joint Institute of Nuclear Research, ul. Joliot-Curie 6, Dubna, Moscow region, 141980, Russian Federation
E-mail: abrekhov@lcc.chem.msu.ru

Abstract. Enzymatic reactions play key role in metabolism of all living organisms. Recent advances in computational methods allow detailed study of enzyme-catalyzed processes, providing data on all intermediate states along reaction path. Furthermore, experimental data on the structure of the reaction intermediates become more available due to advancing of ultrafast X-Ray spectroscopy. These data describes electronic density of intermediates, sometimes even in time series. Currently, prominent Protein Data Bank (PDB) keeps records on the raw crystallographic data on electron density, but usually describes only enzymes or enzymes-inhibitor complexes. A database service where the both experimental and simulated data can be stored and analyzed simultaneously will be a valuable scientific tool for bringing theory and experiment together.

There are hundreds of enzymes known and one should expect large volumes of data on three-dimensional property like electronic density. At the same time, quantum chemistry simulations of large biomolecules require extensive computing resources. Thus, we turned into service model using Nautilus environment of Cognitive Hardware And Software Ecosystem Community Infrastructure (CHASE-CI) project as a blueprint. In that environment, each participating organization can donate their hardware resources, like storage or compute elements, while retaining flexible access control to running services and datasets. We have installed and tested a small-scale environment prototype using resources of LCC laboratory at Chemistry Department Lomonosov Moscow State University, Joint Institute of Nuclear Research (Dubna) and RSC Group company (Moscow). As it was efficiently implemented in Sloan Digital Sky Survey (SDSS) project, we identified basic search queries and structured database accordingly.

1. Introduction
In many fields of science the amount of accumulated data becomes enormous. The problem of effective storage and analysis of these data nowadays is becoming more and more important.
One of such research fields is study of protein-catalyzed reactions. The development of the ultrafast X-Ray spectroscopy allows to look in details on the intermediate reaction steps with extremely high time resolution. The resulting data are time series of three-dimensional electron density maps along the reaction path, i.e. the data are four-dimensional. One may understand the mechanism of the reaction by comparing the experimental electron density maps with those obtained in computer simulations. Today the amount of the experimental data is not too large and such analysis can be done manually. However, it will increase enormously in next years. Thus, the problem of its automatic analysis becomes more and more important.

The first and the most important problem is data storage. The ideal storage system should be large, extensible and reliable. Also, it is important to provide rapid access to the data for analysis. A prominent solution is to use cloud-base storage. In current study we tested distributed cloud-based solution deployed on the regular HPC servers, personal computers and even virtual machines. The reliability of data is provided by partial replication among the nodes. The environment is orchestrated by Kubernetes system.

2. The view on the service from the user’s perspective
The design of the system assumes, that it would be used by the consortium of multiple organizations. Each member will provide the computational and storage resources for shared use. The resources can be geographically distributed and connected through the internet. Depending on needs of each member they will have different level of access to the shared resources. Currently the system is developed within the one organization, but we expect more when it comes to production.

Prior to developing of the system architecture it is important to identify probable use-cases to avoid possible failures at later steps. For example the developers of the Sloan Sky Digital Survey system recommend [1] to estimate around 20 requests. These requests can be obtained for example by surveying the future users. According to these recommendations we planned the design of the database which will be used to store 3D maps metadata.

The database will be filled in a (semi-)automatic way. For example, the computed data on the reaction intermediates will be published in the database after the successful simulation. The metadata will be used to search through the database. An example (incomplete) list of user requests are:

- Find all molecular structures available by the protein name, class, protein structure codes in Protein Data Bank (PDB);
- Find all intermediates and all known 3D maps (if any) by reaction type using enzyme code (EC) classification;

Next, we are going to implement indexing of 3D electron density maps basing on their similarity. There are few commonly used similarity metrics used to compare 3D data. Despite the fact, that the general problem of 3D data comparison is very complex, proteins related to the same enzyme class are alike each other. It allows to use relatively simple similarity metrics and comparison algorithms. For example, 3D electron density distributions are usually compared by constructing differential density map. To compute differential map, the input maps need to be properly aligned in 3D space. If needed, they are normalized and interpolated on the same 3D grid. Next, at each 3D grid point a difference between input maps is computed. This approach is commonly used in ultrafast X-Ray spectroscopy. It is planned to provide such analysis tools to users as they are needed.

3. Platform description
The platform comprises following services:

- Kubernetes system to orchestrate Docker containers.
• Distributed file system based on Ceph object storage. The object storage will be used to store large datasets like 3D maps.
• Database for metadata storage. We are going to use free open-source PostgreSQL database.
• High performance computing services. They will be used to run demanding calculations by user request.
• Web server for user access.

The structured metadata is stored directly in database tables. It includes the data which will be used in search queries, such as protein name, type, intermediate/transition state, etc. The metadata includes links to other databases, for example PDB codes. Semi structured data is stored in the JSON fields in the same database. The database also contain links to the bulky data in the object storage. Examples of these objects are density maps, calculation logs, protein structures.

4. Testing
To demonstrate the flexibility of the distributed platform we used three-node system, using the high-performance server provided by the RSC group, a desktop computed provided by the Department of Chemistry, and the virtual machine provided by the Joint Institution of Nuclear Research in Dubna, connected through the internet. To check the capabilities of this distributed system we used a series of tests for image processing. The dataset of 25,000 images was put to Ceph storage. To model the production workload we used computationally intensive job of image classification using ResNet network [2]. This test has I/O pattern similar to the production use. As a result, almost no performance degradation was observed when using this geographically distributed nodes setup.

5. Conclusion
This work is a part of a project aiming for developing an informational system for analysis of elementary steps of enzyme catalyzed reactions, including analysis of the electronic density. Using modern open-source cloud-based solutions it is possible to configure a geographically distributed informational system in a short time. The system features flexible sharing of the computational and storage resources between multiple users in different organizations. It can be used not only to store the data, but also to manage high-performance calculations. We plan to use the system as a tool for analysing enzyme reactions mechanisms.

Acknowledgement
This work is supported by Russian Foundation for Basic Research (project #18-29-13006).

References
[1] Szalay A.S., Gray J., Thakar A.R., Kunszt P.Z., Malik T., Raddick J., Stoughton C., vandenBerg J. The SDSS skyserver: public access to the sloan digital sky server data. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (SIGMOD ’02). ACM, New York, NY, USA, 570-581. DOI: https://doi.org/10.1145/564691.564758.
[2] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. CoRR, abs/1512.03385. 2015.