Contribution ID: **111** Type: **Presentations**

BRUSELAS: A HPC based software architecture for drug discovery on large molecular databases

Thursday, 30 November 2017 14:45 (15 minutes)

1. Overview

In the context of computer-aided drug discovery (CADD), virtual screening (VS) is a collection of in-silico techniques to filter large molecular databases searching for bioactive compounds. Such techniques, together with the development of high performance computing (HPC) infrastructures, allow the access to a huge number of compounds in a short time and at a low cost.1

In the last years, a diversity of VS web servers and HPC platforms has been developed.2 However, there is still a gap to be covered: the need of an architecture capable of scalably integrating a large number of data sources and extract a drug candidate list that fits the user needs. Aiming at filling this gap, we have developed BRUSELAS (Balanced Rapid and Unrestricted Server for Extensive Ligand-Aimed Screening) which is a software architecture to perform 3D similarity searches on large datasets of compounds using HPC techniques.

BRUSELAS exhibits a modular design capable of importing data coming from several sources containing very diverse contents. It is accessible free of cost at http://bio-hpc.ucam.edu/Bruselas, and its development is based on the experience acquired in previously awarded DECI / PRACE projects.

2. BRUSELAS and Big Data

A recurrent question is to what extent molecular databases represent Big Data resources. In scientific literature, they are usually considered as Big Data because they accomplish the 5Vs rule in the following terms:3
1. Volume - represented by the number of compounds in the existing databases.
2. Variety - determined by the collection of entities and properties of very different nature stored in such databases.
3. Velocity - given by the time employed to screen datasets.
4. Veracity - validated by comparing the predicted results with experimental ones.
5. Value - given by the success of the process.

In cases where similarity algorithms handle flexibility, a set of conformers representing diverse poses are generated for each compound. In such situations, the volume is given by the total number of conformers to screen, which usually is quite larger than the initial amount of compounds. In addition, the comparison of millions of conformers is a very slow process which can be significantly accelerated by using HPC resources.

3. Intended audience

VS is a topic of growing interest that is closely related to bioinformatics, biochemistry, HPC and Big Data. Furthermore, it is of potential application in many fields such as theoretical and experimental chemistry as well as in more applied biology and medicine areas.

References

1. Bajorath J. (2002) Integration of virtual and high-throughput screening. Nat. Rev. Drug. Discov. 1(11):882–894
2. Pérez-Sánchez H., Rezaei V., Mezhuyev V., Man D., Peña-García J., Den-Haan H., Gesing S. (2016) Developing science gateways for drug discovery in a grid environment. Springerplus. 5(1):1300
3. Laney D. (2001) 3D Data management: controlling data volume, velocity and variety. Appl. Deliv. Strateg. Internet(February 2001):1–4

Topic Area

The EOSC & EDI building blocks

Type of abstract

Presentation (15 minutes)

Primary author: BANEGAS-LUNA, Antonio-Jesus (Universidad Católica San Antonio de Murcia)

Co-authors: Dr CABALLERO, Alberto (Universidad Católica San Antonio de Murcia); Dr PÉREZ-SÁNCHEZ, Horacio (Universidad Católica San Antonio de Murcia)

Presenter: BANEGAS-LUNA, Antonio-Jesus (Universidad Católica San Antonio de Murcia)

Session Classification: EOSC building block presentations