E-Biothon: an experimental platform for BioInformatics

Michel Daydé, Benjamin Depardon, Alain Franc, Jean-François Gibrat, Romaric Guillier, Yasaman Karami, Christian Pérez, Frédéric Suter, Marie Chabbert, Bruck Taddese, et al.

E-Biothon: an experimental platform for BioInformatics

Michel Daydé
CNRS - IRIT, Université de Toulouse / INPT-ENSEEIHT
Toulouse, France
e-mail: michel.dayde@cnrs-dir.fr

Benjamin Depardon
SysFera
Lyon, France
e-mail: benjamin.depardon@sysfera.com

Alain Franc
INRA - UMR BioGeCo, Cestas, France
Inria - Pleiade Team, Talence, France
e-mail: alain.franc@inria.fr

Jean-François Gibrat
INRA, UR1404, Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement, 78350 Jouy-en-Josas, France
e-mail: jean-francois.gibrat@france-bioinformatique.fr

Romaric Guilllier
CNRS - LIP ENS Lyon
Lyon, France
e-mail: romaric.guillier@ens-lyon.fr

Yasaman Karami
Unité de Biologie Computationnelle et Quantitative, UMR 7238 CNRS-UPMC, Paris, France
e-mail: yasaman.karami@upmc.fr

Christian Pérez
AVALON Team, INRIA / LIP, ENS Lyon
Lyon, France
e-mail: christian.perez@inria.fr

Frédéric Suter
IN2P3 Computing Center, CNRS/IN2P3, LIP ENS Lyon
Lyon, France
e-mail: frederic.suter@cc.in2p3.fr

Bruck Taddese, Marie Chabbert
Laboratoire BNMI, UMR CNRS 6214 INSERM U1083, Faculté de médecine, Angers, France
e-mail: marie.chabbert@univ-angers.fr

Sylvie Thérond
CNRS - IDRIS, Orsay, France
e-mail: sylvie.therond@idris.fr

ABSTRACT

The E-Biothon platform [8] is an experimental Cloud platform to help speed up and advance research in biology, health and environment. It is based on a Blue Gene/P system and a web portal that allow members of the bioinformatics community to easily launch their scientific applications. We describe in this paper the technical capacities of the platform, the different applications supported and finally a set of user experiences on the platform.

Keywords

Bioinformatics, parallel computing, Cloud.

1. INTRODUCTION

France has always been at the forefront of biology, health and environment researches. When considering the major "epidemics" and pathologies of our time such as AIDS, cancer, diabetes, the genetic and proteomic analysis of pathogenic agents or of patients plays an increasing role in helping to develop new treatments. Recent technological advances, such as high-throughput sequencers, enable biology researchers to have access to massive amounts of raw data (petabytes of data are generated every year) on the composition of viruses, bacteria, plants and animals (including the human species). Analyzing this data to take advantage of the information available is a crucial task that requires a very large amount of computer processing capacity. As a consequence, parallel computing is becoming more and more central in advancing the research in life sciences.

In order to provide parallel computing resources and help the life science community to prepare their codes for the next generation of massively parallel computers, CNRS, IBM, Inria, the Institut Français de Bioinformatique and SysFera have joined forces to give researchers access to the E-Biothon Cloud platform, hosted at IDRIS [13], the CNRS major centre for High Performance Computing that is also one of the three national supercomputing centres in France, located in Orsay, near Paris. Combining an application portal with very large computing power, it will make it possible to develop software and applications that will spur on research in biology and health, in particular in genomics, proteomics and metabolomics. The objective is to speed up research in several areas e.g. agronomy, marine biology, biotechnologies, fundamental biology or to improve knowledge of genetic diseases as rapidly as possi-

ble, particularly neuromuscular diseases, and to drastically speed up the discovery of new breakthrough treatments. The platform also aims to accelerate research in ecology-biodiversity in order to enhance our understanding of the environment.

Since using distributed computing systems is not always easy for people that do not have a computer science background, E-Biothon is designed to provide a simple access to powerful tools enabling their research.

2. PLATFORM DESCRIPTION

The platform comprises four racks of the high performance IBM Blue Gene/P systems, representing a power of 56 teraflops associated with 200 terabytes of storage, and the SysFera-DS solution, which offers users a web portal for accessing the computational resources. Through this portal, researchers have access to a complete working environment allowing them to easily carry out computerized processing related to analysis in the fields of genomics, proteomics and metabolomics, and then manage the data generated, all through a simple web browser.

SysFera-DS communicates with the available batch scheduler (in this case LoadLeveler) to submit jobs on behalf of the users and track them throughout their lifecycle. The web interface provides a simple way for the users to input the data corresponding to a selected number of parameters required by the applications and to transfer back their results, rather than going through the command-line or SSH transfers.

3. APPLICATIONS

This section presents the applications that are currently available on the platform. It is not a definitive list as we are always eager to support new applications, depending on the requirements of our community of users. It also does not account for the applications that we did not integrate in the portal that are using the computing resources of the platform.

3.1 PhyML

PhyML [10] is a phylogeny software based on the maximum-likelihood principle. Early PhyML versions used a fast algorithm performing Nearest Neighbor Interchanges (NNIs) to improve a reasonable starting tree topology. Since the original publication [11] in 2003, PhyML has been widely used (more than 10,000 citations on Google Scholar), because of its simplicity and a fair compromise between accuracy and speed. It is still being developed by the LIRMM in Montpellier, which also provides an online computing facility through their website [2] that is mostly dedicated to bootstrapping jobs. The E-Biothon platform completes this setup by providing an environment for longer jobs that require more computing power.

It is currently the most used application on the platform accounting for 95% of the registered users and about 55% of the monthly computing time used (note that in the past, other applications corresponding to today finished projects strongly occupied the platform).

3.2 NAMD

NAMD [23] is a parallel molecular dynamics (MD) code designed for high-performance simulation of large biomolecular systems, originally developed by the University of Illinois. It is designed to be able to scale out over a large number of cores (over 500,000) and be able to process complex structures at atomic-level detail, such as the HIV capsid that contains more than 1,300 proteins and 64 million atoms. As a widely used software in the field of molecular dynamics (more than 4,000 references to their 2005 publication), it currently represents 40% of the monthly computing time used on the E-Biothon platform.

3.3 LAMMPS

LAMMPS [24] is a massively parallel simulation tool for the movement of molecules, developed by the Sandia National Laboratories. It is designed to efficiently compute Newtons equations of motion for collections of atoms, molecules, or macroscopic particles that interact via short- or long-range forces with a variety of initial and/or boundary conditions.

4. USER TESTIMONIES

A significant percentage of the users of the platform are from France but practically they come from more than 30 different countries (as seen in Figure 1) and work in a large range of fields from molecular dynamics to comparative genomics. This section aims at presenting some of their experiences and results achieved through the E-Biothon platform.

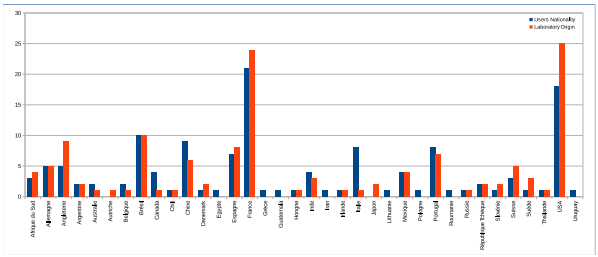

Figure 1: Origin of the E-Biothon users and their research laboratories

4.1 Chemokine Receptor CXCR4

The aim of the project is the detailed analysis of the dynamics properties of the CXC chemokine receptor 4 (CXCR4). This receptor belongs to the G protein-coupled receptor family (GPCRs). It is involved in a wide variety of diseases, including tumor cell metastasis, autoimmune and inflammatory diseases in addition to serving as a co-receptor for HIV-1 viral entry [3]. In 2010, Wu et al. [27] reported the first crystal structures of CXCR4 co-crystallized with small molecules that give a static view of the inactive state. However, proteins are inherently dynamic systems and GPCRs can be described in terms of conformational ensembles [6]. The receptors sample diverse distinct conformations that lead to different downstream functions and are influenced by binding of different ligands. Knowledge of these conformations and of reactional paths between them will help develop drugs that selectively prevent or induce specific downstream functions.

Accelerated MD [20] has been shown to be an efficient way to enhance conformational sampling and to reduce the computational time necessary to observe major activation/deactivation conformational changes by several orders of magnitude. Using this method, we analyze the conformational ensembles sampled by

CXCR4 upon different conditions (active or inactive state, mutations, etc.). Figure 2 shows the increased conformational sampling obtained by accelerated MD compared to classical MD. In either case, the simulations started from an inactive conformation of CXCR4. Two markers (the TM3-TM6 distance and the RMSD of the NPxxY motif) are representative of conformational states. We also analyze the transitions between different states with special emphasis on time-dependent correlated conformational changes of side chains surrounding functional GPCR microswitch residues that trigger the activation/deactivation transition [7, 22]. MD simulations are run using NAMD software [23]. The system contains ∼63,000 atoms for monomeric receptor, POPC lipids, ions and TIP3 water molecules. Using E-Biothon with 512 CPU jobs, each nanosecond of trajectory requires about 2h30. The web interface makes submission and data management user-friendly.

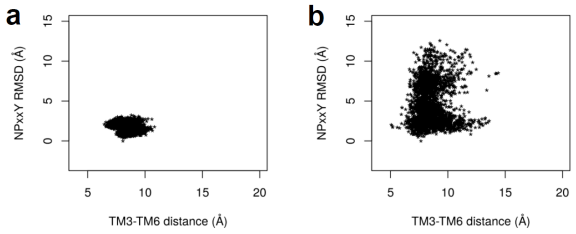

Figure 2: Conformational ensembles sampled by CXCR4 during a 60 ns trajectory obtained by classical (a) and accelerated (b) MD simulations at 310 K with NAMD.

4.2 Biodiversiton

This project is a use of massive parallelization for biodiversity studies. Genomes are imprints of past evolution, and their variability permits to corroborate, or not, the organization of diversity of living organisms as built by studies in Natural History. This is the Barcode of Life project. Nowadays, NGS technologies produce millions of reads (short sequences) per environmental sample (a community). Sequences can be compared to a reference database by local alignment scores [21].

Making all comparisons necessitates (for one data set) 3.6×10^{12} elementary operations. Only massive parallelization on a Blue Gene could permit such brute force calculation. As we have at the same time the morphological based inventory of the same communities (10 in Leman Lake, 19 in Swedish rivers), it is the first time that an exact comparison is made between a morphological based inventory and the annotation of a metagenome from a database for the taxonomic composition of a community. Usual technique for such inventories is a massive use of BLAST, but parallelization of BLAST is not easy.

Hence, we have written a code in C, called `disseq`, implementing Needleman-Wunsch algorithm for sequence comparison[18], and the loop for running over all pairs has been parallelized with MPI, in a program called `mpidisseq`, as a collaboration between the research group in BioGeCo and IDRIS. This first experiment (all comparisons have been produced, and post-treatment is currently ongoing) is followed by a deeper investigation on pairwise distances between reads within an environmental sample, in a project accepted by DARI (french procedure for applying to resources on the national computing centres managed by GENCI[9]) on Turing (which is one of the national supercomputers in France located at IDRIS), with massive parallelization currently running on a BlueGene Q.

4.3 COMMA

We have developed a tool, Communication Mapping (COMMA), that identifies the dynamical architecture of proteins from all-atom molecular dynamics (MD) simulations in explicit solvent [17]. Growth Hormone (GH) is a four helix bundle that regulates a wide variety of physiological processes, including growth and differentiation of muscle, bone, and cartilage cells [25]. The regulation of normal human growth is initiated by the binding of GH to its receptor (GHR), where the complex of GH-GHR is a homodimer (a 1:2 complex in which GH binds to two identical subunits of the receptor) [5].

The GH-GHR system consists of around 100,000 atoms. We studied the wild type (WT) and one pathogenic mutant (MU) of this protein (R183H) [28]. We used NAMD to produce the trajectories [23] and performed 2 replicates of 100-ns MD simulations for the WT and MU, to detect networks of dynamically correlated residues. The all-atom root mean square deviation (RMSD) from the equilibrated structure were recorded along each 100-ns MD simulation replicate (Figure 3), for the WT and MU (R183H) complexes, to assess their stability. We used an estimation of 90,000 CPU hours for the MD simulations, on BlueGene/P machine in HPC mode. The obtained trajectories serve as the input to COMMA.

COMMA defines *communication blocks*, that are groups of residues with high communication propensity and strong non-covalent interactions and maps this information on the structure of the protein. We applied COMMA on GH-GHR to extract communication blocks of WT and MU (Figure 4). The set of 8 communication blocks in WT (pink, red, brown, yellow, dark pink, orange, sand and magenta) and 3 in MU (pink, orange and red) are shown on the cartoon representation of the structure. Pathways that correspond to the differences between blocks are colored in green. The schematic representations of the largest blocks (> 60 residues) in WT and MU are depicted on the left. The results indicate a dynamics-based rewiring of communication network in GH-GHR induced by the deleterious mutation (manuscript in preparation). COMMA provides hints on how the mutation affects the dissociation of the GH-GHR and enables us to detect the key pathways on the structures of the WT and MU.

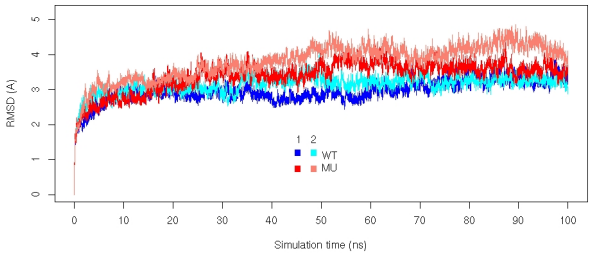

Figure 3: RMSD for the Growth Hormone Complex. Colors correspond to the two replicates (1 and 2) of the WT and MU.

4.4 Insyght

Insyght[19] is a tool to carry out bacterial comparative genomics. The software, available through a web-

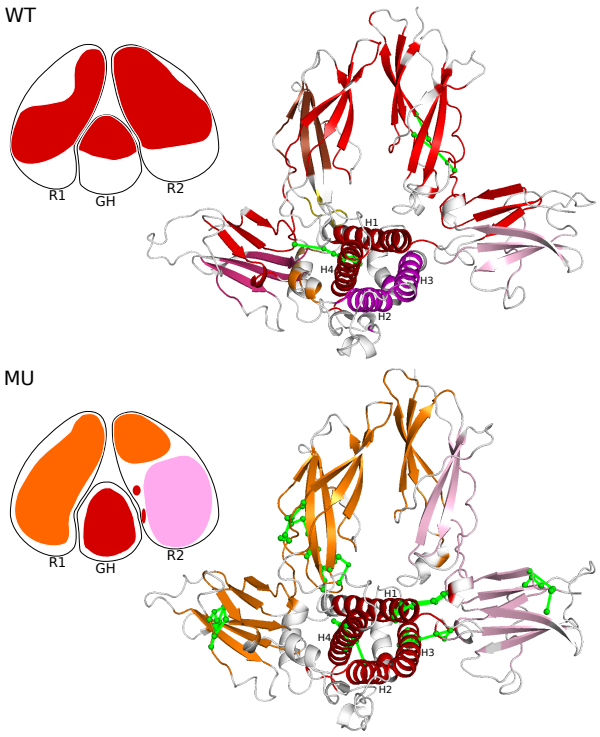

Figure 4: Comparison between Communication blocks of GH-GHR in WT and MU (R183H).

site[16], tightly integrates three complementary views: (i) a table for browsing among homologs, (ii) a comparator of ortholog functional annotations and (iii) a genomic organization view that combines symbolic and proportional graphical paradigms to improve the legibility of genomic rearrangements and distinctive loci. Insyght benefits from an easy and smooth navigation between these 3 views and provides users with a powerful search mechanism. Data for performing these comparisons are precomputed and stored in a relational database.

The computation of these data requires (i) the cross comparison with BLASTp [1] of the proteomes (i.e., all the proteins coded by a genome) and (ii) the determination of syntenie regions for all the pairs of bacterial genomes stored in the relational database. Here, 2,660 complete bacterial genomes from Ensembl Bacteria were used, giving rise to more than 3.5 millions pairs of bacterial genomes. BLASTp jobs generated 1.2 TB of raw, compressed data. The final relational database has a total size of 3.5 TB, the largest table having about 5.9 billions lines and occupying ∼1 TB of disk space (∼2 TB with indexes).

This work would not have been possible without the E-Biothon computing power as the number of jobs required is quadratic with the number of genomes (> 3.5 millions) and the users did not have enough computer resources locally to execute the months' worth of computations it represented.

5. CONCLUSION

The E-Biothon platform is open to the life science community. It is still evolving for incorporating new applications. The success of this project is not only coming from the availability of a parallel computing platform since the user support provided by all the partners of the project (CNRS, IBM, IDRIS, Inria, Institut Français de Bioinformatique and SysFera) is crucial for efficiently deploying new applications and managing the platform.

6. ACKNOWLEDGEMENT

The authors would like to thank CNRS [4], IBM [12], IDRIS [13], Inria [15], l'Institut Français de Bioinformatique [14] and SysFera [26] without which it would not have been possible to setup the project. We are also very grateful to all the people that have been experimenting the platform since the beginning of the project.

7. ADDITIONAL AUTHORS

REFERENCES

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 1990.

[2] ATGC montpellier phyml online execution. `http://www.atgc-montpellier.fr/phyml/`.

[3] W.-T. Choi, S. Duggineni, Y. Xu, Z. Huang, and J. An. Drug discovery research targeting the CXC chemokine receptor 4 (CXCR4). *J Med Chem*, 2011.

[4] CNRS website. `http://www.cnrs.fr/index.php`.

[5] A. M. de Vos, M. Ultsch, and A. A. Kossiakoff. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, 255(5042):306–312, Jan 1992.

[6] X. Deupi and B. Kobilka. Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. *Physiology*, 2010.

[7] X. Deupi and J. Standfuss. Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr Opin Struct Biol*, 2010.

[8] E-Biothon portal. `https://www.ebiothon.fr`.

[9] GENCI: Grand Equipement National de Calcul Intensif web site. `http://www.genci.fr`.

[10] S. Guindon, J. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, 2010.

[11] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 2003.

[12] IBM website. `http://www.ibm.com`.

[13] IDRIS website. `http://www.idris.fr/`.

[14] Institut Français de Bioinformatique web site. `http://www.france-bioinformatique.fr`.

[15] Inria website. `http://www.inria.fr/`.

[16] INSYGHT website. `http://genome.jouy.inra.fr/Insyght/`.

[17] Y. Karami, E. Laine, and A. Carbone. Dissecting protein architecture with communication blocks and communicating segment pairs. *BMC bioinformatics*, 2015.

[18] L. Kermarrec, A. Franc, F. Rimet, P. Chaumeil, J.-M. Frigerio, J.-F. Humbert, and A. Bouchez. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33:349–363, 2014.

[19] T. Lacroix, V. Loux, A. Gendrault, M. Hoebeke, and J. Gibrat. Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol! *Nucleic Acids Research*, 2014.

[20] P. R. L. Markwick and J. A. McCammon. Studying functional dynamics in bio-molecules using accelerated molecular dynamics. *Phys. Chem. Chem. Phys.*, 2011.

[21] S. B. Needleman and C. D. Wunsch. A general method applicable to search for similarities in the amino-acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[22] R. Nygaard, T. Frimurer, B. Holst, M. Rosenkilde, and T. Schwartz. Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol Sci*, 2009.

[23] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kal, and K. Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 2005.

[24] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 1995.

[25] M. Sundstrom, T. Lundqvist, J. Rodin, L. B. Giebel, D. Milligan, and G. Norstedt. Crystal structure of an antagonist mutant of human growth hormone, G120R, in complex with its receptor at 2.9 A resolution. *J. Biol. Chem.*, 271(50):32197–32203, Dec 1996.

[26] Sysfera website. `https://www.sysfera.com/`.

[27] B. Wu, E. Chien, C. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. Bi, D. Hamel, P. Kuhn, T. Handel, V. Cherezov, and R. Stevens. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science*, 2010.

[28] Y. L. Zhu, B. Conway-Campbell, M. J. Waters, and P. S. Dannies. Prolonged retention after aggregation into secretory granules of human R183H-growth hormone (GH), a mutant that causes autosomal dominant GH deficiency type II. *Endocrinology*, 143(11):4243–4248, Nov 2002.