ATGC et E-Biothon

Romaric GUILLIER[1], Frédéric SUTER[2], Benjamin DEPARDON[3], Vincent LEFORT[4]

[1]CNRS - LIP ENS Lyon
romaric.guillier@ens-lyon.fr

[2]CNRS CC-IN2P3, LIP ENS Lyon
frederic.suter@cc.in2p3.fr

[3]SysFera
benjamin.depardon@sysfera.com

[4]CNRS - LIRMM, Université de Montpellier
Vincent.Lefort@lirmm.fr

Deux plates-formes

ATGC: Origine

- Projet ATGC Montpellier (CNRS, UM, IFB, LIRMM)
  - Collaboration entre les laboratoires IBC, IRMB, LGDP, LIRMM, MiVEGEC et South Green
  - Orientée vers la génomique évolutive, comparative et fonctionnelle
  - ATGC est labellisée par IBiSA et coordonne les activités de services en bioinformatique sur le Sud de la France pour l'Institut Français de Bioinformatique (IFB)
  - Buts:
    - Diffuser et valoriser les outils bioinformatiques développés au sein de la communauté régionale
    - Favoriser les collaborations entre partenaires informaticiens et biologistes
    - Apporter une aide à ces chercheurs en mettant en place des services bioinformatiques en lien direct avec leurs travaux

E-Biothon: Origine

- 2001-2012: projet Decrypthon[BBC+10] (CNRS, IBM, AFM)
  - Accélérer la recherche sur la génétique et les maladies rares
  - Donner un accès transparent à des ressources distribuées de calcul
  - Grille de calcul sur 6 sites (Bordeaux, Lille, ENS Lyon, Paris 6, Orsay, UPMC), DIET middleware et un portail web.
- 2013-201X: projet E-Biothon[DGG+15] (CNRS, IBM, Inria, IFB, SysFera)
  - Contexte: Les séquenceurs génétiques sont capables de générer beaucoup de données, mais nécessitent beaucoup de capacité de calcul et des logiciels efficaces.
  - Buts:
    - Aider les chercheurs en sciences du vivant à mettre au point des codes adaptés pour des machines massivement parallèles.
    - Donner un accès facile à des applications existantes.

ATGC: Description

- Machine (hébergée par le LIRMM)
  - Cluster de 16 noeuds de calcul PowerEdge R630 bi-pro Intel Xeon E5-2650 v3, 2,3GHz => 320 cores
  - 192 GB RAM par noeud => 3072 GB
  - 30 TB de stockage
- Portail web (hébergé par le LIRMM)
  - Applications disponibles: CompPhy, FastME, MPScan, MS ALIGN, PhyML, PhySIC_IST, PhySIC, RED2, ReplacementMatrix, STAR
  - Soumission libre via l'interface web
  - Suivi et envoi des résultats par email
  - Tâches les plus gourmandes en temps de calcul rejetées pour éviter la saturation du service. Ces utilisateurs sont invités à utiliser E-Biothon pour les tâches facilement parallélisables pouvant bénéficier de l'architecture de la BlueGene.

E-Biothon: Description

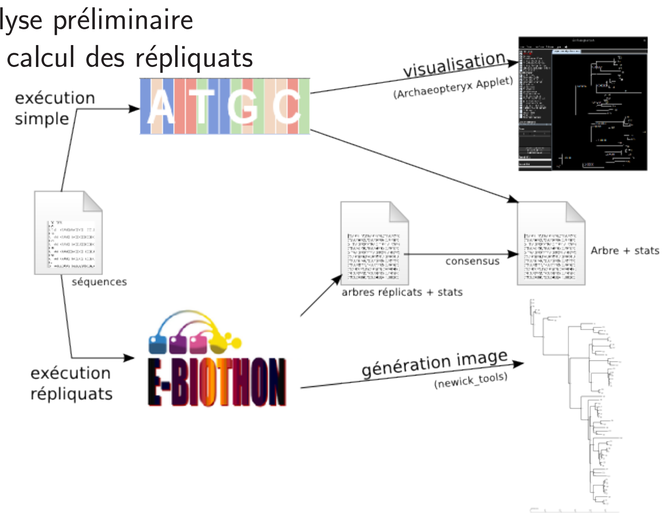
- Machine (hébergée par l'IDRIS)
  - IBM BlueGene/P
  - 56 Tflops de performance pic
  - 4 racks de 1024 CPU (PPC32) de 4 cores chacun => 16384 cores
  - 2 GB RAM par CPU => 8192 GB
  - 200 TB de stockage
  - réservation max: 1024 CPU * 120 h
- Portail web (hébergé par le CNRS)
  - Applications disponibles: PhyML, NAMD, LAMMPS, GROMACS
  - Accès sur inscription (limité aux chercheurs institutionnels)
  - SysFera-DS pour gérer les réservations
  - Suivi et récupération des résultats sur le site, notification par email.

Approche similaire pour simplifier le travail des chercheurs en bio-informatique

Une application phare: PhyML

Logiciel

- Développé au LIRMM[GDL+10, GG03], cité plus de 10000 fois
- Construction d'arbres phylogénétiques optimisant le critère maximum de vraisemblance
- Nombre d'arbres possibles en $O(\mathbf{T}^{\mathbf{T}})$
- Complexité calcul en $O(\mathbf{I} * \mathbf{T}^2)$ ($\mathbf{T}$, nombre de séquences; $\mathbf{I}$, longueur des séquences)
- Besoin d'exécuter nombreuses réplications de calcul ("bootstrap") pour obtenir un indicateur statistique sur les branches de l'arbre résultat.
- Complémentarité
  - ATGC: analyse préliminaire
  - E-Biothon: calcul des répliquats



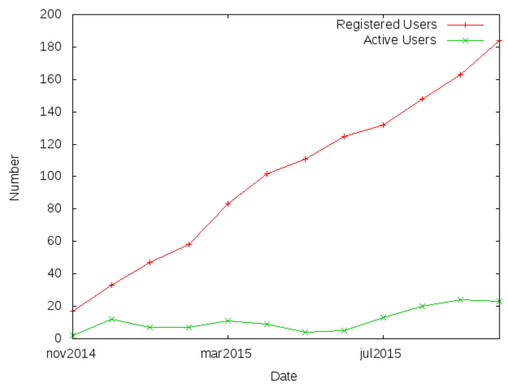
Usage

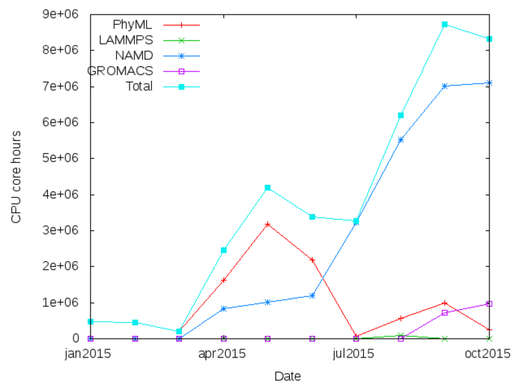
ATGC

- 3000 jobs soumis mensuels (d'utilisateurs du monde entier)

E-Biothon

- 95% des utilisateurs inscrits pour PhyML (sur 200 utilisateurs de 40 pays différents)
- 30-50% des jobs mensuels soumis (sur 250 jobs mensuels)
- 10-70% du volume de calcul





Références

[BBC+10] N. Bard, R. Bolze, E. Caron, F. Desprez, M. Heymann, A. Friedrich, L. Moulinier, N.H. Nguyen, O. Poch, and T. Toursel, *Décrypthon grid - grid resources dedicated to neuromuscular disorders*, Stud. Health Technol. Inform. (2010).

[DGG+15] Michel Daydé, Jean-François Gibrat, Romaric Guillier, Bruck Taddese, Marie Chabbert, Alain Franc, Yasaman Karami, Sylvie Thérond, Benjamin Depardon, Frédéric Sutter, Christian Pérez, and Gilles Lesage, *E-biothon: an experimental platform for bioinformatics*, CSIT (2015).

[GDL+10] S. Guindon, J.F. 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).

[GG03] S. Guindon and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*, Systematic Biology (2003).

Partenaires

Email: e-biothon@services.cnrs.fr
vincent.lefort@lirmm.fr

WWW: https://www.e-biothon.fr
http://www.atgc-montpellier.fr