



Inria

IBM



Historique et Infrastructure de E-Biothon : une plateforme expérimentale pour la bioinformatique

M. Daydé,

CNRS – IRIT / Université de Toulouse / INPT

Basé sur des présentations précédentes avec pour co-auteurs :

A. Cheniour, Ph. Collinet, B. Depardon, F Desprez, A. Franc, JF Gibrat, R Guillier, Y Karami, C Pérez, F Suter, B Taddese, M Chabbert, S Thérond

Etapes préliminaires

Téléthon 2001 :

- AFM et IBM appel à la mobilisation des internautes : « Mettez à la disposition de la recherche la puissance inutilisée de votre PC ».
- Objectif : réaliser la première cartographie du protéome : l'ensemble des protéines/molécules produites par les cellules.
- Succès : 75 000 internautes mobilisés, des milliards de calculs complexes effectués, 550 000 protéines du monde vivant cartographiées.
- Production d'une véritable bibliothèque de comparaison des protéines des différentes espèces du monde vivant (animal, végétal, humain) avec près de 2,2 millions de fichiers répartis en 17 000 répertoires.
- Chaque ordinateur a contribué pour environ 133 heures, i.e. 10 millions d'heures de calcul au total. 21 serveurs IBM ont hébergé solutions + données sur les 2 mois de l'opération.

Appel à projets en 2003 pour exploiter cette BD avec 4 projets retenus :

- Etude des relations entre la structure et les fonctions des protéines qui réduisent les risques d'anomalies génétiques chez l'homme et la levure.
- Identification et caractérisation de protéines impliquées dans plusieurs maladies neuromusculaires, ainsi que prédiction des domaines protéiques et des fonctions tissu-spécifiques
- Analyse des protéines du peroxysome, impliqué dans de nombreuses fonctions métaboliques essentielles
- Identification de nouvelles cibles thérapeutiques potentielles chez *Vibrio cholerae* et les organismes Diabac (Diarrhoeal Bacteria) impliqués dans les pathologies diarrhéiques.

2 projets de plus sélectionnés en 2003/2004

- But : démontrer la faisabilité d'un programme disposant de sa propre grille de calcul pour le mettre à la disposition de toutes les équipes
- Projets portés avec succès sur la grille qui en bénéficient pour leurs calculs.

Projet DECRYPTHON

2004-2012

- Suite au succès des opérations précédentes
- Convention est signée en mai 2004 entre l'AFM, le CNRS et IBM, officialisant le projet Décrypthon
- Objectifs
 - Accélérer la recherche sur les maladies génétiques et les maladies rares
 - Rendre transparent l'utilisation de ressources de calcul distribuées aux utilisateurs
- Infrastructure :
 - Une grille de calcul constituée de 6 sites (Bordeaux, Lille, ENS Lyon, Paris 6, Orsay and UPMC)
 - Intergiciel DIET et portail Web
 - Retrait de l'AFM en 2012



Contexte du projet E-Biothon



- Suite à des discussions démarrées en 2012 à l'arrêt du Décryphon CNRS, IBM, Inria, l'Institut français de Bioinformatique et SysFera s'entendent pour le déploiement de E-Biothon.
- Permet de mettre au point les logiciels et les applications qui permettront d'accélérer la recherche en biologie et en santé, en particulier en génomique et en protéomique mais aussi en écologie-biodiversité afin de mieux comprendre notre environnement
- Annonce officielle du lancement à SC'13 avec 3 applications notamment dans les domaines de l'épidémiologie et de la bio-diversité.

Plateforme E-Biothon et outils

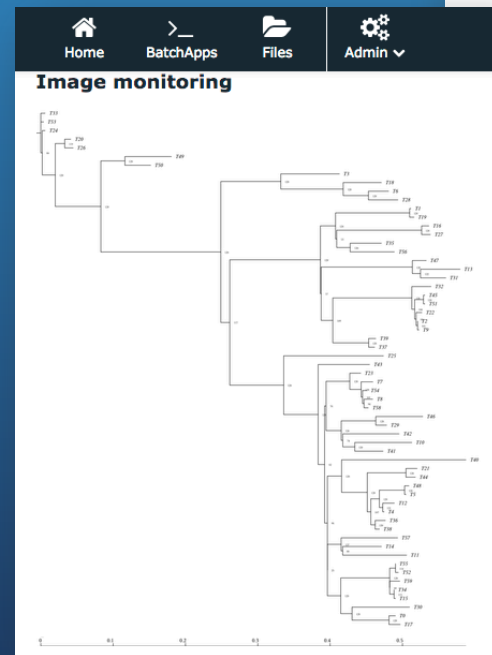
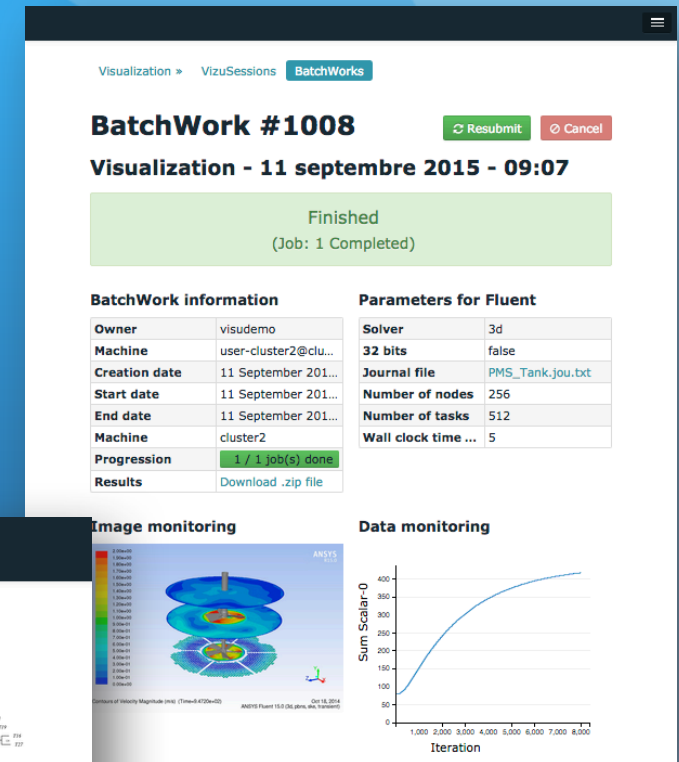
Calculateur de E-Biothon : BlueGene/P hébergé à l'**IDRIS**

- 4 racks de BLueGene/P (Ex-Babel) hébergés à partir de septembre 2014
 - Après reconfiguration et recâblage de la machine initialement de 2 racks
 - Performance crête 56 Tflops
 - Chaque rack à 1024 nœuds de 4 cœurs
 - 200 To de stockage
- Deux modes d'exploitation : standard (HPC) or High Throughput Computing (HTC) avec quelques limitations



SysFera-DS Intuitive Web Interface

- Brings HPC & Cloud environment to local desktop
- Runs non-interactive and interactive graphical HPC applications
- Manages & visualizes remote Big Data
- Manages projects & access rights
- Follows application & resource usage



Job Submission

- Submission form specific per application
- Single or multi-step jobs
- Output preview

namd13 - 31 juillet 2015 - 11:46

Finished
(Jobs: 3 Completed)

BatchWork information

Owner	rguillier
Machine	phym101@babel
Creation date	31 July 2015 - 11:46:42
Start date	31 July 2015 - 11:48:34
End date	31 July 2015 - 12:00:29
Machine	babel
Submission script	view
Progression	3 / 3 job(s) done

Parameters for namd13-job-multistep

Input Tarball	test.tar
Time Limit	10
CPU Number	1024
Initial Step	0
Last Step	2

There are 3 jobs attached to this BatchWork

Completed

Job #2544

Completed

Job #2545

Completed

Job #2546

New BatchWork (step 2)

Selected application: PhyML

Input Data

Sequence file *  Please select a sequence file

Data type *  ☒ DNA ☐ Amino-acids

Sequence file type *  interleaved

Number of datasets *  1

Substitution model

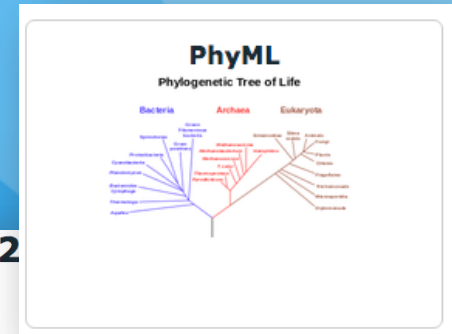
Substitution model * GTR

Equilibrium frequencies *  ☐ optimized ☒ empirical

Transition/transversion ratio  0,00 Fixed Estimated

Proportion of invariable sites  0,00 Fixed Estimated

Number of substitution rate categories *  4



Remote file management

Users can access their remote files on Babel

- Upload/download
- Preview results
- Copy/move/delete
- Rights

The screenshot displays the GROMACS BatchWorks interface. At the top, there's a navigation bar with 'GROMACS', 'BatchWorks', 'Activity Report', 'Statistics', and 'Settings'. Below this, a 'BatchWork #2162' section shows a progress bar for zipping a file (358 out of 1647, 32.8 MB / 75.7 MB) and buttons for 'Download output', 'Resubmit', and 'Cancel'. A green status bar indicates 'Finished (Job: 1 Completed)'. The main area is titled 'Machines' and shows a 'Browse' button. Below this is a table of files for machine 'babel'. A context menu is open over the file 'YLIIP.txt_phymI_boot_stats.txt', showing options like 'View/Edit', 'Download', 'Rename', 'Delete', 'Change group', 'Change permissions', and 'Close'. The table has columns for Name, Last modification, File size, Owner, Group, and Permissions. At the bottom, there's a 'Transfer list' table with columns for 'source file' and 'destination folder'.

GROMACS - 16 September 2015 - 15:53

BatchWork #2162 [Download output] [Resubmit] [Cancel]

Ziping file 358 out of 1647 (32.8 MB / 75.7 MB).

Please wait, the archive file is being generated...

Finished
(Job: 1 Completed)

Machines

Browse [View/Edit] [Delete] [Change group] [Change permissions]

Machine : babel

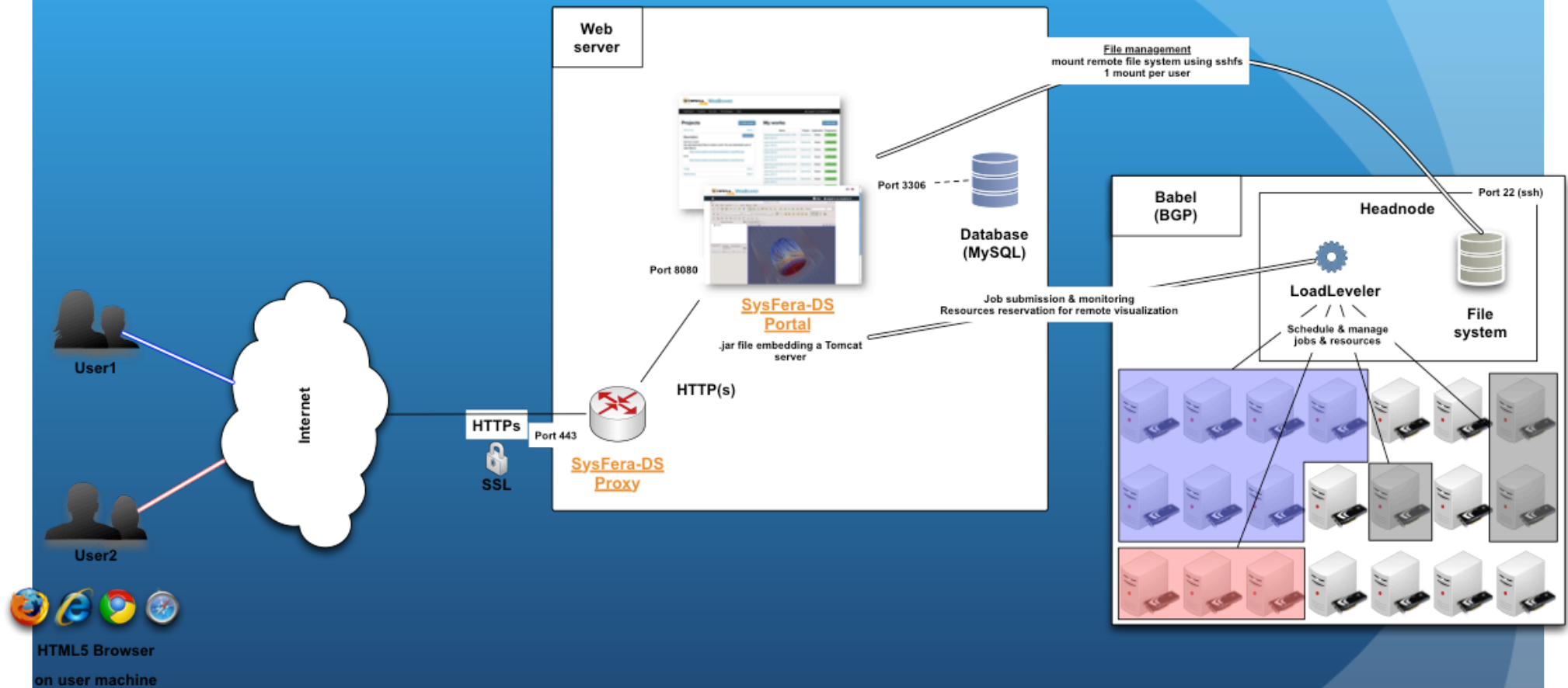
Name	Last modification	File size	Owner	Group	Permissions
..	-	-	-	-	-
clan2musclecuratedPH YLIIP.txt	2015-08-18 11:11:21	13.5 KiB	2000006	2000001	rwxrwxrwx
clan2musclecuratedPH YLIIP.txt_phymI_boot_ stats.txt			2000006	2000001	rwxrwxrwx
clan2musclecuratedPH YLIIP.txt_phymI_boot_ trees.txt			2000006	2000001	rwxrwxrwx
clan2musclecuratedPH					

Displaying 1-7 of 7 | Per page
Order is ascending Name
Current directory: /workgpf/

Transfer list

source file	destination folder
-------------	--------------------

Architecture

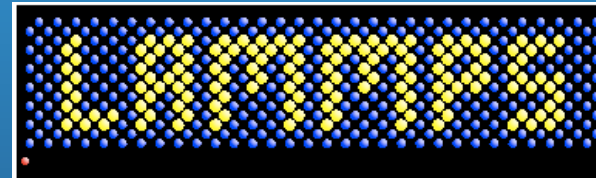


Applications déployées

Applications



- Code parallèle de dynamique moléculaire pour la simulation de systèmes biomoléculaires de grande taille développé initialement à « University of Illinois »
- Conçu pour passer à l'échelle sur un grand nombre de nœuds (plus de 500,000) et pour traiter des structures complexes à l'échelle atomique comme le capsid du HIV qui contient plus de 1,300 protéines and 64 millions d'atomes
- Très utilisé en dynamique moléculaire, représente plus de 40% de l'occupation mensuelle de E-Biothon



- Outil de simulation massivement parallèle pour le mouvement des molécules développé par « Sandia National Laboratories »
- Conçu pour calculer efficacement les équations de mouvement de Newton pour des ensemble d'atomes, de molécules ou de particules macroscopiques interagissant avec des forces à petite ou grande échelle avec diverses conditions initiales ou aux frontières

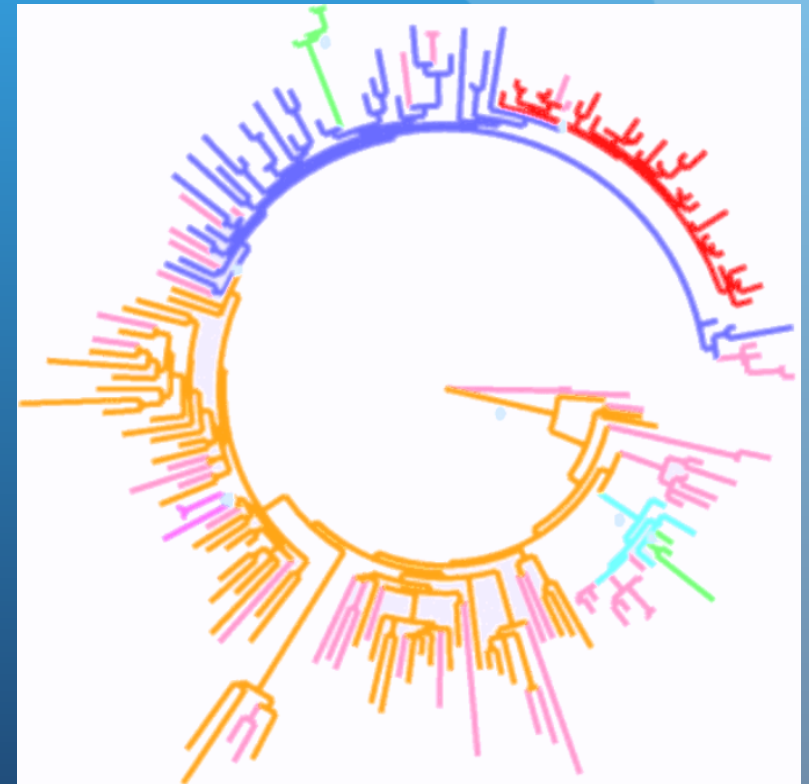


- GROMACS : autre outil de dynamique moléculaire pour des systèmes avec des centaines de millions de particules initialement développé à Groningen University,

PhyML

A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood (ML)

- Inputs :
 - Data : a file containing DNA or protein sequences.
 - An evolutionary model (chosen by the user).
- Principle :
 - Compute the probability to observe the data given the evolutionary model and a phylogenetic tree.
- Complexity problem :
 - For n sequences, the number of possible trees is $O(n^n)$.
 - Cannot compute the likelihood for all trees.
- PhyML algorithm :
 - Based on a hill-climbing approach, which modifies the tree so as to maximize the likelihood.
- PhyML citations :
 - PhyML paper is the most cited in ecology-environment since 2007 (see Science Watch).
 - It was cited more than 10 000 times.
- Some application fields :
 - Genomics : to predict gene function and identify therapeutic targets.
 - Medicine : to trace the origin of epidemics and prevent new pandemics.
 - Ecology : to inventory the biodiversity and to preserve environment.

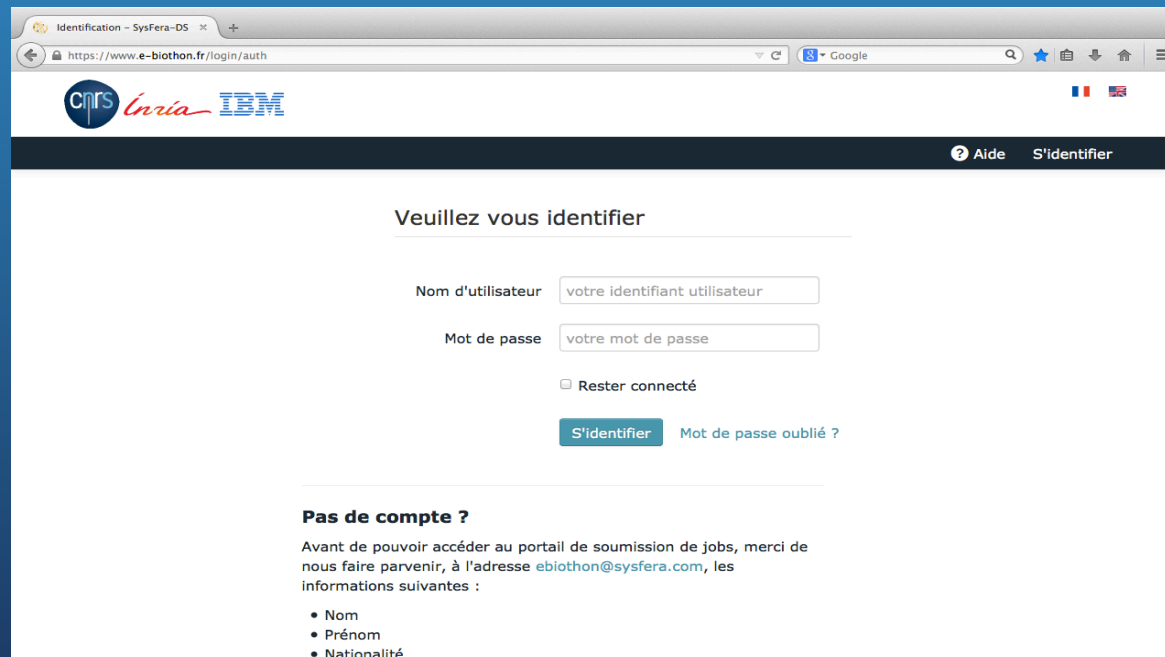


Projet PhyML

- Depuis octobre 2014, possibilité pour utilisateurs de PhyML d'employer la plate-forme E-Biothon lorsque leur demande ne peut pas être satisfaite par la plate-forme ATGC : <http://www.atgc-montpellier.fr/phyml/>
- Soumission de travaux via le portail SysFera-DS, à travers une interface dédiée réalisée par SysFera
- Procédure d'agrément des utilisateurs mise en place (déclaration nominative et acceptation de la charte informatique du CNRS) : une quinzaine de demandes en novembre 2014
- Modalités d'accès
 - Demande par mail à ebiothon@sysfera.com faite depuis l'adresse académique (laboratoire ou organisation) pour accord
 - Utilisateur reçoit son login et mot de passe provisoire pour pouvoir se connecter au Portail Sysfera-DS

Page Authentification

- Portail Sysfera-DS accessible à l'url suivante:
<https://www.e-biothon.fr/login/auth>
- 1^{ère} Connexion avec login et mot de passe provisoire pour le nouvel utilisateur.
- Personnalisation du mot de passe.



The screenshot shows a web browser window with the address bar displaying <https://www.e-biothon.fr/login/auth>. The page header includes logos for CNRS, Inria, and IBM, along with a language selector (French and English flags) and links for 'Aide' and 'S'identifier'. The main content area is titled 'Veuillez vous identifier' and contains two input fields: 'Nom d'utilisateur' with the placeholder 'votre identifiant utilisateur' and 'Mot de passe' with the placeholder 'votre mot de passe'. Below these fields is a checkbox labeled 'Rester connecté' and a blue 'S'identifier' button. A link 'Mot de passe oublié ?' is positioned to the right of the button. At the bottom, there is a section titled 'Pas de compte ?' with a paragraph explaining that users need to send an email to ebiothon@sysfera.com with their personal information: Nom, Prénom, and Nationalité.

Identification - SysFera-DS

<https://www.e-biothon.fr/login/auth>

CNRS Inria IBM

Aide S'identifier

Veuillez vous identifier

Nom d'utilisateur

Mot de passe

☐ Rester connecté

[Mot de passe oublié ?](#)

Pas de compte ?

Avant de pouvoir accéder au portail de soumission de jobs, merci de nous faire parvenir, à l'adresse ebiothon@sysfera.com, les informations suivantes :

- Nom
- Prénom
- Nationalité

Page Accueil

The screenshot shows a web browser window titled "PhyML - SysFera-DS - Iceweasel". The address bar shows the URL "https://www.e-biothon.fr/project/show/2". The page features a header with logos for CNRS, Inria, and IBM, and a navigation bar with links for "Projets", "Mes tâches", and "Gestionnaire de fichiers". A user is logged in as "user". The main content area is titled "PhyML" and includes sections for "Description", "Membres", "Dernières expériences", "Quotas", and "Applications". Each section contains a message indicating that the user does not have the necessary permissions to view the information. A "Créer une expérience" button is visible in the "Dernières expériences" section. The footer displays the SysFera-DS logo, the version "RELEASE_V_5.13-102-g4fb7efc", and the copyright notice "Copyright © SysFera 2011-2014 — Tous droits réservés".

PhyML - SysFera-DS - Iceweasel

https://www.e-biothon.fr/project/show/2

PhyML - SysFera-DS - Iceweasel

Projets Mes tâches Gestionnaire de fichiers Aide Connecté(e) en tant que user

PhyML > Expériences Créer une expérience

PhyML

Description

PhyML

Membres

Vous ne disposez pas des permissions nécessaires pour voir cette information.

Dernières expériences

Vous ne disposez pas des permissions nécessaires pour voir cette information.

+ Créer une expérience

Quotas

Vous ne disposez pas des permissions nécessaires pour voir cette information.

Applications

PhyML (batch)

A test version of the PhyML application.

Version RELEASE_V_5.13-102-g4fb7efc

Copyright © SysFera 2011-2014 — Tous droits réservés

Créer une expérience 1/3

https://www.e-biothon.fr/work/create/2?execution=e1s1

cnrs *Inria* IBM

Projets ▼ Mes tâches Gestionnaire de fichiers Aide Connecté(e) en tant que user ▼

PhyML > Expériences Créer une expérience

Nouvelle expérience (étape 1 sur 2)

Sujet

Application

Machine cible



Créer une expérience 2/3

https://www.e-biothon.fr/work/create/2?execution=e1s2

cnrs Inria IBM

Projets ▼ Mes tâches Gestionnaire de fichiers ? Aide Connecté(e) en tant que user ▼

Nouvelle expérience (étape 2 sur 2)

Input Data

Sequence file * ? nucleic_M2573_346x897_2006.phy (remove)

Data type * ? ☒ DNA ☐ Amino-acids

Sequence file type * ? interleaved

Number of datasets * ? 1

Substitution model

Substitution model * GTR

Equilibrium frequencies * ? ☐ optimized ☒ empirical

Transition/transversion ratio ? 0 Fixed Estimated

Proportion of invariable sites ? 0 Fixed Estimated

Number of substitution rate categories * ? 4

Gamma shape value ? 0 Fixed Estimated

Créer une expérience 3/3

The screenshot shows a web browser window with the URL <https://www.e-biothon.fr/work/create/2?execution=e1s2>. The page is titled "Tree searching" and contains several configuration sections:

- Starting Tree**: Includes a "Please select a tree file" section with a "File" button and a "BIONJ" button.
- Tree improvements type**: A dropdown menu set to "NNI".
- Number of random starting trees**: A numeric input field set to "0".
- Optimize topology ***: Radio buttons for "yes" (selected) and "no".
- Optimize branch lengths ***: Radio buttons for "yes" (selected) and "no".
- Branch Support**: Includes a "Bootstrap *" input field set to "100" and a "Bootstrap CPU number *" dropdown set to "50 (2 task(s)/r)". A note below states: "64 nodes will be reserved (14 nodes unused). To optimize your reservation, you could set your bootstrap to 128."
- Scheduling options**: Includes a "Wallclock limit *" input field set to "3".

At the bottom right, there are two buttons: "Retour" (red) and "Valider" (green). The footer of the page displays the "SYSPERA 7DS" logo and the text "Version RELEASE V 5.13-102-q4fb7efc".

Suivi du statut

https://www.e-biothon.fr/job/show/1661

Google

cnrs Inria IBM

Projets ▼ Mes tâches Gestionnaire de fichiers Aide Connecté(e) en tant que user ▼

PhyML » Expériences Créer une expérience

Tâche #J_1661

Créée par PhyML - 25 novembre 2014 - 12:07

Résultats de la tâche

Statut	COMPLETED
Identifiant de la tâche	J_1661
Date de soumission	2014-11-25 12:07:46 CET
Date de fin	2014-11-26 04:04:01 CET
Identifiant de la machine	babel
Nom d'hôte/IP	babel.idris.fr
Identifiant de la tâche ...	babel1-adm.idris.fr.375435.0
Nom de la tâche pour l'...	PhyML
File d'attente	Mrt3
Groupe	ebiothon
Fichier stdout	babel.idris.fr:/home/pfs/idris/ebiotho...
Fichier stderr	babel.idris.fr:/home/pfs/idris/ebiotho...
Répertoire de sortie	/workgpf/idris/ebiothon/phyml01/us...

Paramètres de soumission

Propriétaire	phyml01
Session	webboard-session-31512
Priorité	Immediate
Répertoire de travail	/workgpf/idris/ebiothon/phyml01/user
Temps d'exécution	71 940
Nombre de nœuds	1
Nombre de processeur...	1:0

SYSFERA OS

Résultats

https://www.e-biothon.fr/fileBrowser/index?machine=babel&folder=/workgpfs/idris/ebiotion/phyml01/user/VISHNU_OUTPU

Google

cnrs Inria IBM

Projets ▼ Mes tâches Gestionnaire de fichiers Aide Connecté(e) en tant que user ▼

Machines

Actions sur les fichiers sélectionnés

Sélectionner Voir/Modifier Supprimer Changer le groupe Changer les permissions

Machine : babel

Nom	Dernière modif.	Taille	Propriétaire	Groupe	Permissions
..	-	-	-	-	-
nucleic_M2573_346x89_7_2006.phy	2014-11-25 18:38:24	93 B	phyml01	ebiotion	rw-rw-rw-
nucleic_M2573_346x89_7_2006.phy.phyml_boo	2014-11-26 04:08:19	59.5 KiB	phyml01	ebiotion	rw-rw-rw-
t_stats.txt					
nucleic_M2573_346x89_7_2006.phy.phyml_boo	2014-11-26 04:08:19	1033.3 KiB	phyml01	ebiotion	rw-rw-rw-
t_trees.txt					
nucleic_M2573_346x89					

Affichage de 1-5 sur 5 | Par page : 20 Tout
Classé par ordre croissant de Nom
Répertoire courant : /workgpfs/idris/ebiotion/phyml01

Liste des transferts		
fichier source	dossier de destination	Actions

SYSFERA DS
Version RELEASE V 5.13-102-n4fb7efr

SysFera-DS utilise des cookies pour ses fonctionnalités. En utilisant SysFera-DS, vous acceptez notre utilisation de cookies. OK

Expérimentations

Correlated Motions in Distinctive Conformational States of the Chemokine Receptor CXCR4, B. Taddese and M. Chabbert,

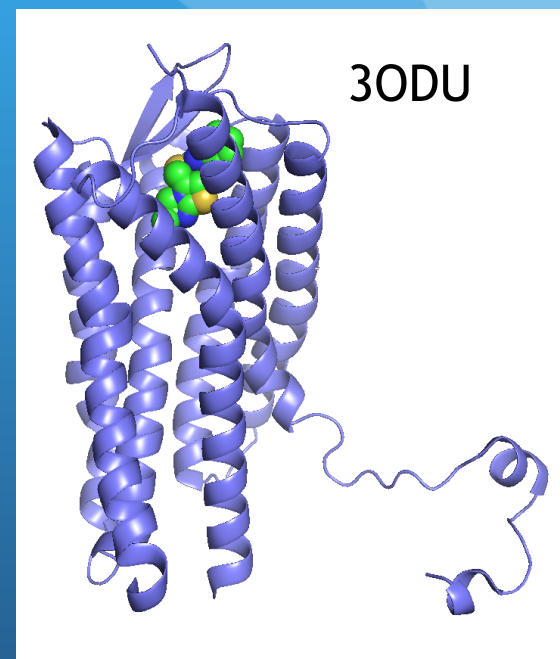
UMR CNRS 6214 - INSERM 1083, University of Angers, France

- Detailed analysis of the dynamics properties of the CXC chemokine receptor 4 (CXCR4)
- Belongs to G protein coupled receptor (GPCR) family
Involved in inflammation, chemotaxis, neural development and cancer and also co-receptor for HIV-1 viral entry
- Crystal structure of inactive CXCR4 studied by Wu et al., 2010

Important drug target

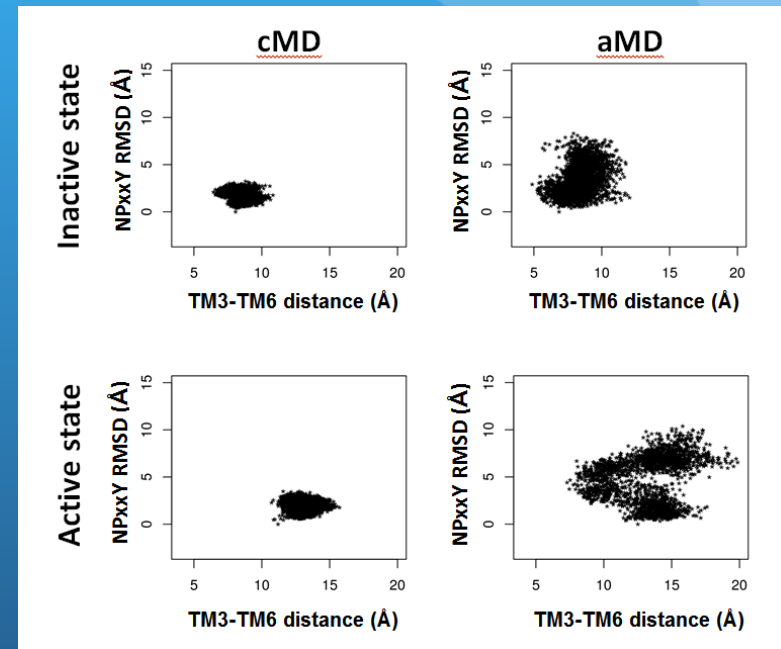
Resolved crystal structure

Prototype of chemotactic receptors



- Dynamics of proteins is important: Molecular dynamics (MD) can only reach nanoseconds to microseconds
- Activation GPCRs on millisecond timescales
- Sampling of conformational space can be improved using **Accelerated Molecular Dynamics** that enhances conformational sampling and reduces the computational time necessary to observe major activation / deactivation conformational changes by several orders of magnitude:
 - Increase computational power
 - algorithmic improvements
 - methodological developments

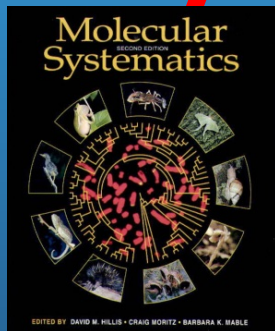
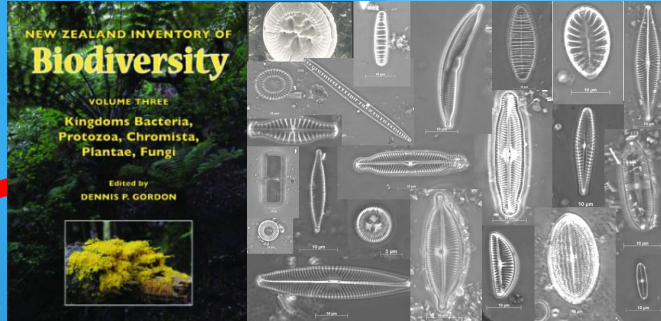
Conformational ensembles sampled by CXCR4 during a 60 ns trajectory obtained by classical (cMD) and accelerated (aMD) MD simulations at 310 K with NAMD.



- System with ~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.

Biodiversiton project

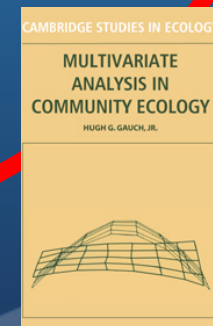
Alain Franc
INRA - UMR BioGeCo /
Inria - Pleiade Team



+

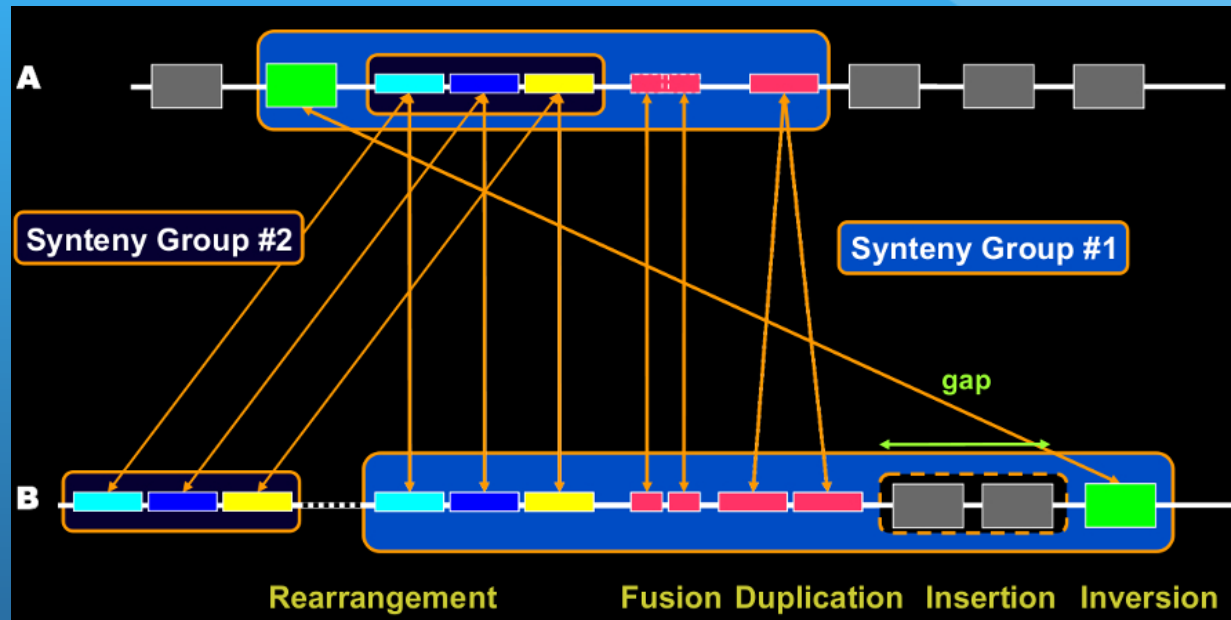


- Utilisation du parallélisme massif pour étudier la diversité biologique et la structure des différentes communautés d'organismes
- Parallélisation des calculs de distance avec MPI et passage sur le DARI (centres nationaux HPC)
- Cf présentation Alain Franc



Insyght: outil de comparaison des génomes procaryotes

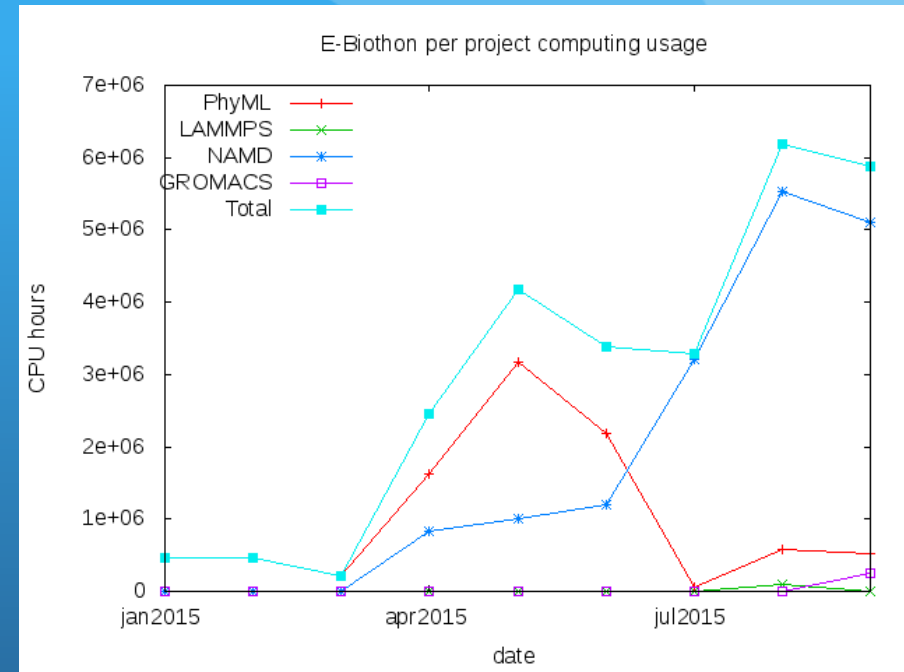
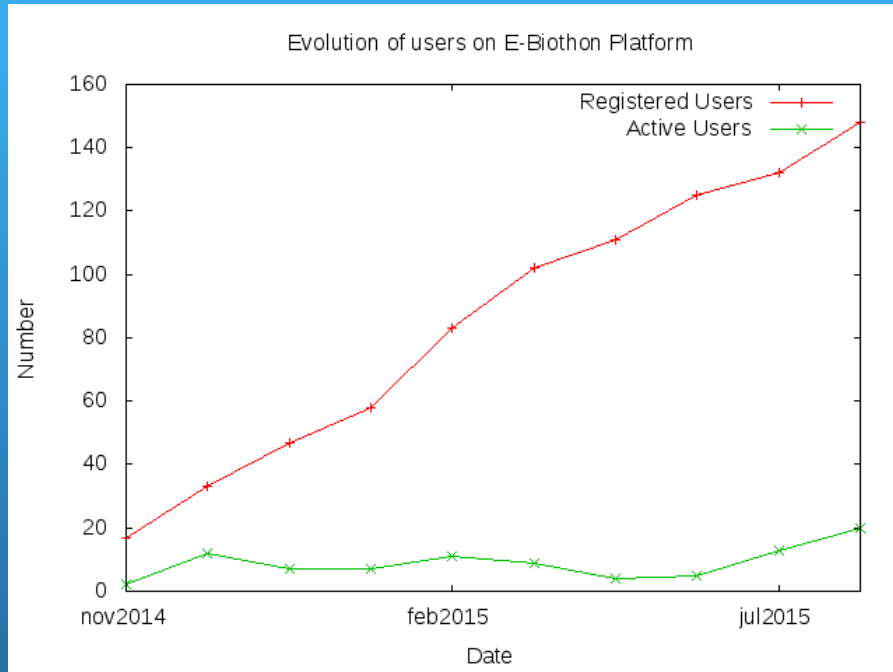
Jean-François Gibrat, INRA, UR1404, Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement



- Comparaison du contenu en gènes de près de 2700 génomes complets de bactéries avec environ 2 mois de la plateforme complète (>3.5 M comparaisons de génomes chacun avec 16 M de gènes, BD de 3.5 To)
- Permet de visualiser les séquences protéiques homologues et les relations de synténie chez les procaryotes pour par exemple étudier les différences génétiques entre des bactéries pathogènes et d'autres non pathogènes
- Insyght facilite cette tâche aux biologistes : accès facile à l'ensemble des données, des résultats, comparaisons de génomes, d'extraction des relations de synténie et visualisation facilement grâce à une interface web accessible à tous.
- Résultats publiés dans la revue Bioinformatics de novembre 2015.

Utilisation de la plateforme

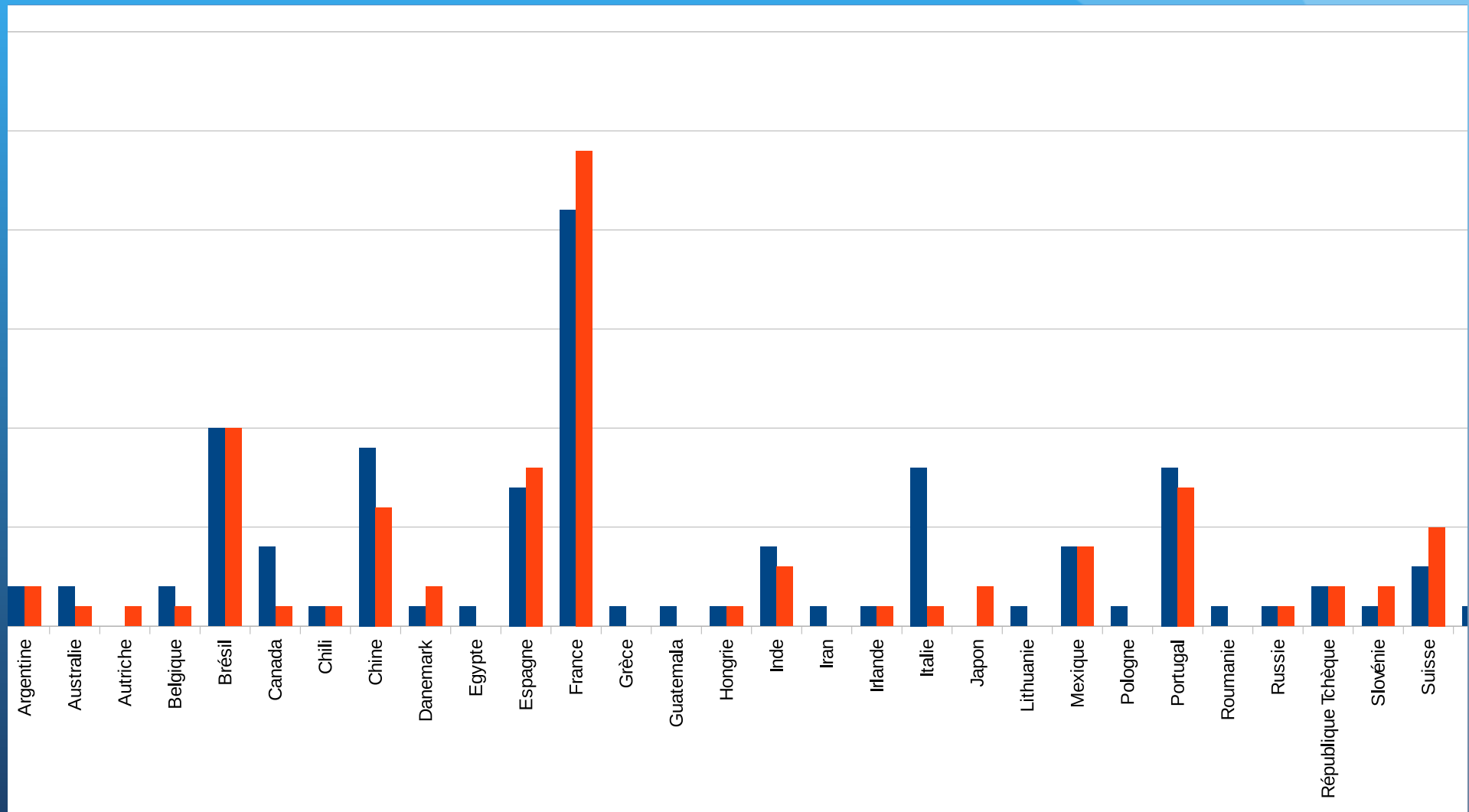
Utilisateurs et charge



Mai 2016 :

- 223 utilisateurs dont 87% pour PhyML
- 88% CPU consacré à NAMD
- CPU Size distribution
 - 1 : 0 (0.00 %)
 - 64 : 77 (64.17 %)
 - 128 : 7 (5.83 %)
 - 256 : 4 (3.33 %)
 - 512 : 14 (11.67 %)
 - 1024 : 18 (15.00 %)

Provenance des utilisateurs





Inria

IBM



Conclusion

- E-Biothon : plateforme pour la communauté des sciences de la vie qui va être arrêtée d'ici peu après 1 an de phase de démarrage et 2 ans d'activité opérationnelle
- Migration des activités vers le DARI et la plateforme IFB récemment installée ?
- Succès du projet pas uniquement lié à disponibilité de la plateforme (Bluegene + interface) :
 - Importance du support déployés par les partenaires (CNRS, IBM, IDRIS, Inria, Institut Français de Bioinformatique and SysFera)
 - Crucial pour déployer de nouvelles applications et gérer la plateforme
- *Remerciements à CNRS, IBM, IDRIS, Inria, l'Institut Français de Bioinformatique and SysFera sans qui ce projet n'aurait pas vu le jour et aux utilisateurs !*