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

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 Décrypthon 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.

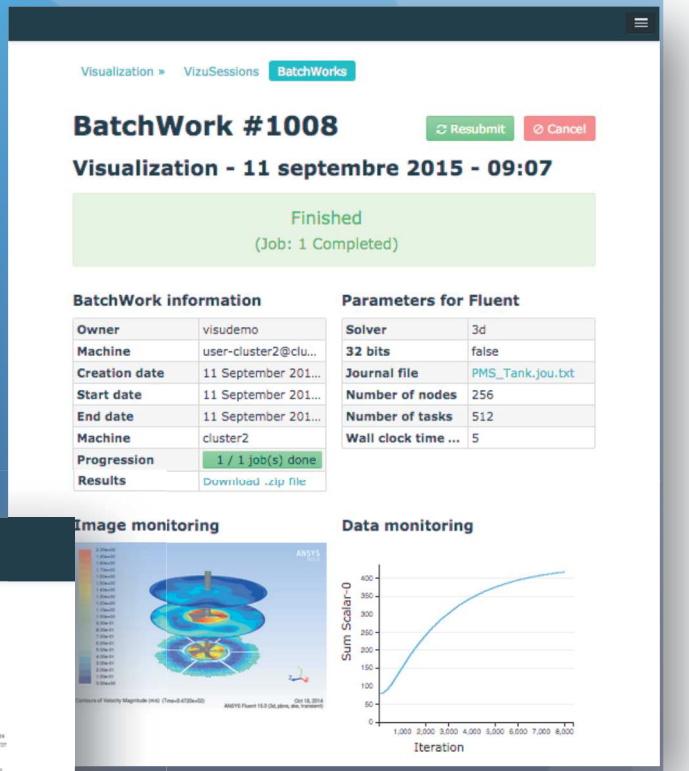
Platforme E-Biothon et outils

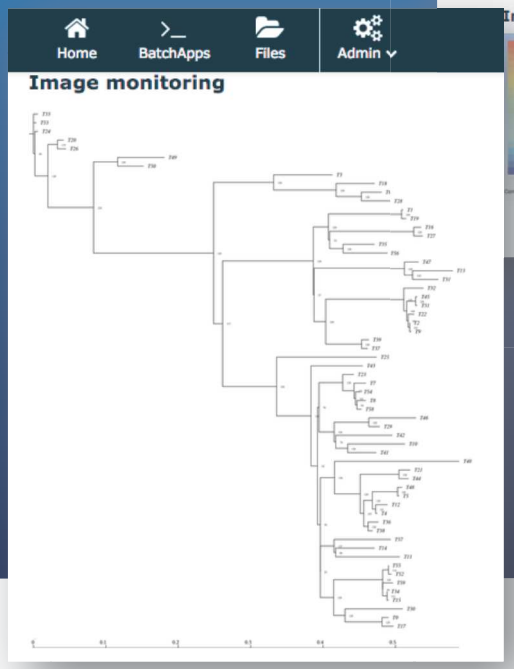
Calculateur de E-Biothon : BlueGene/P hébergé à l'IDRIS

- 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 Throughtput Computing (HTC) avec quelques limitations

SysFera-DS Intuitive Web Interface

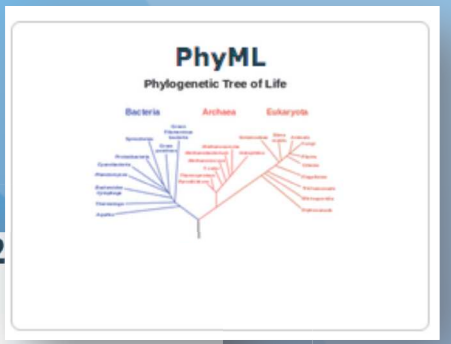
- Brings HPC & Cloud environment to local desktop
- Runs non-interactive and interactive graphical HPC applications
- Manages & visualizes remote Big Data
- Manages projects & access rights
- Follows application & resource usage





Job Submission

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



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

namd13 - 31 juillet 2015 - 11:46

Finished
(Jobs: 3 Completed)

BatchWork information

Owner	rguillier
Machine	phyml01@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

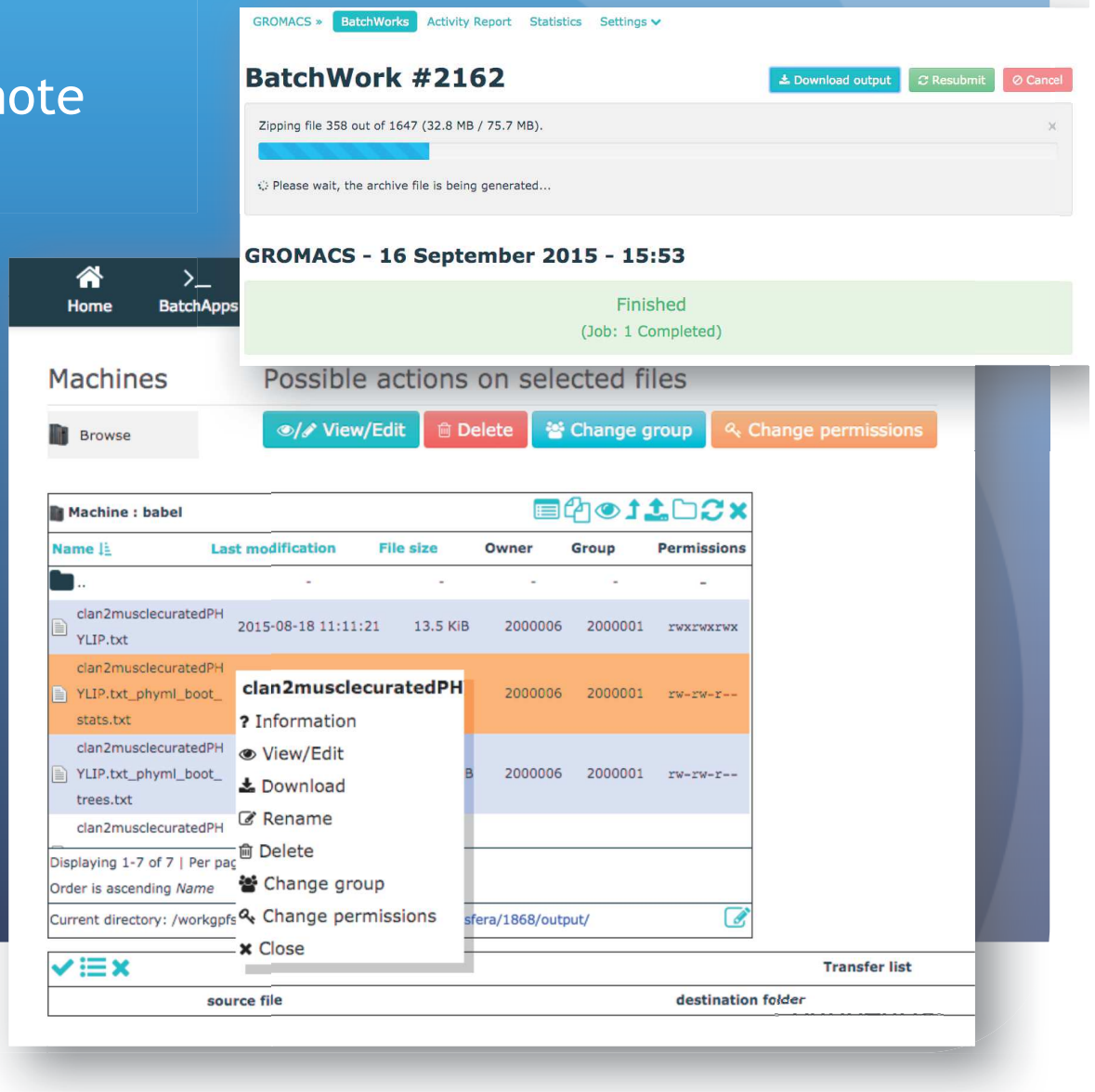
Completed — Job #2545

Completed — Job #2546

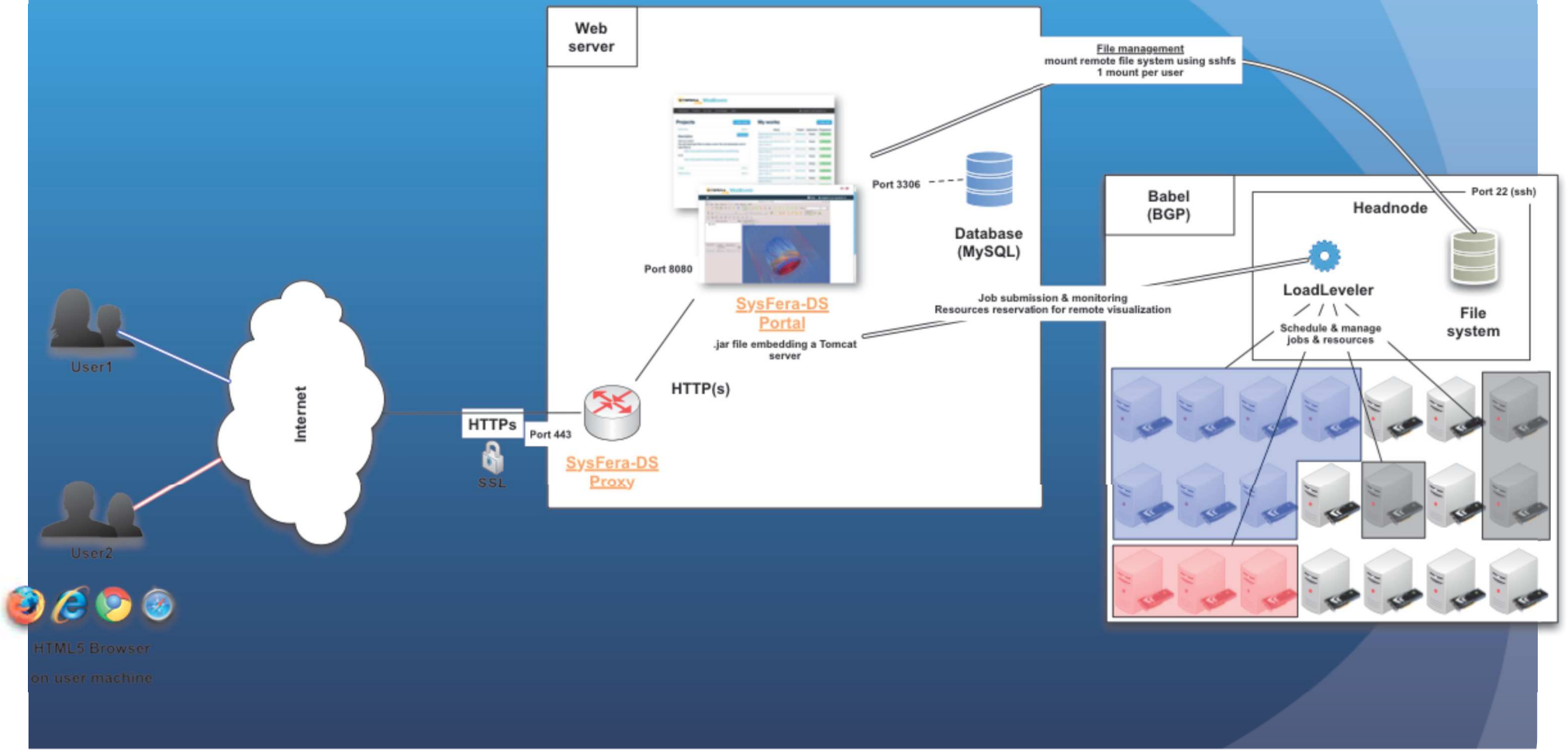
Remote file management

Users can access their remote files on Babel

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



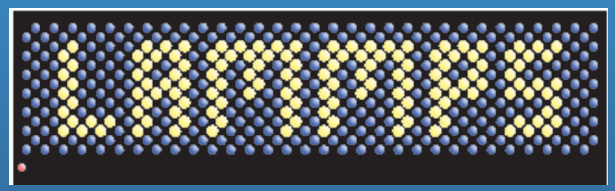
Architecture



Applications déployées

Applications

- Code parallèle de dynamique moléculaire pourla 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 capside 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



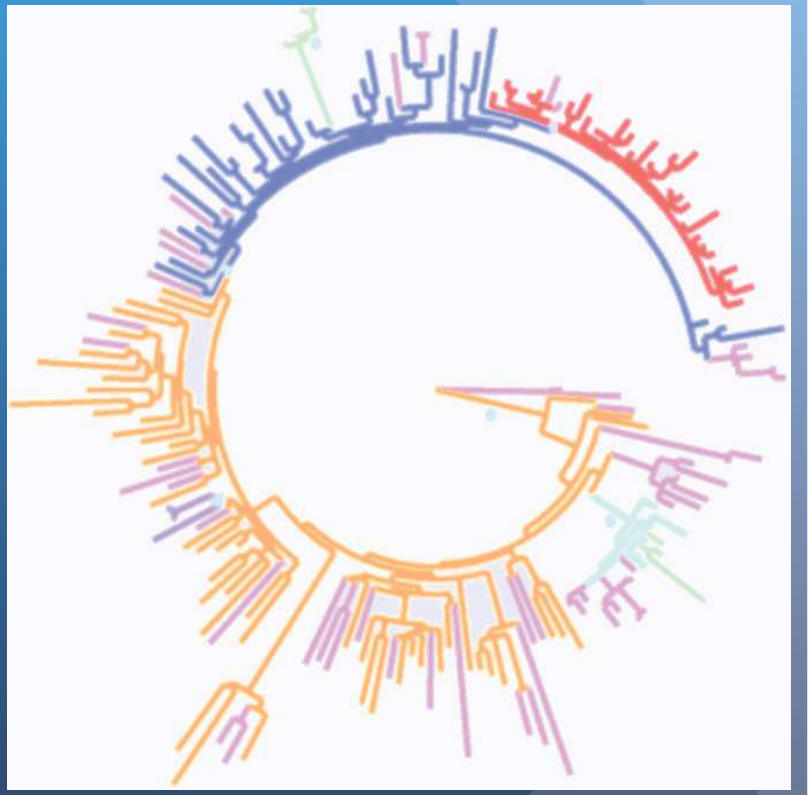
- Outil de simulation massivement parallèle pour le mouvement des molécules développé par « Sandia National Laboratories »
- Conçu pour calculer efficacement les équations de mouvement de Newton pour des ensemble d'atomes, de molécules ou de particules macroscopiques interagissant avec des forces à petite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

PhyML

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

South of France bioinformatics platform

- 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.

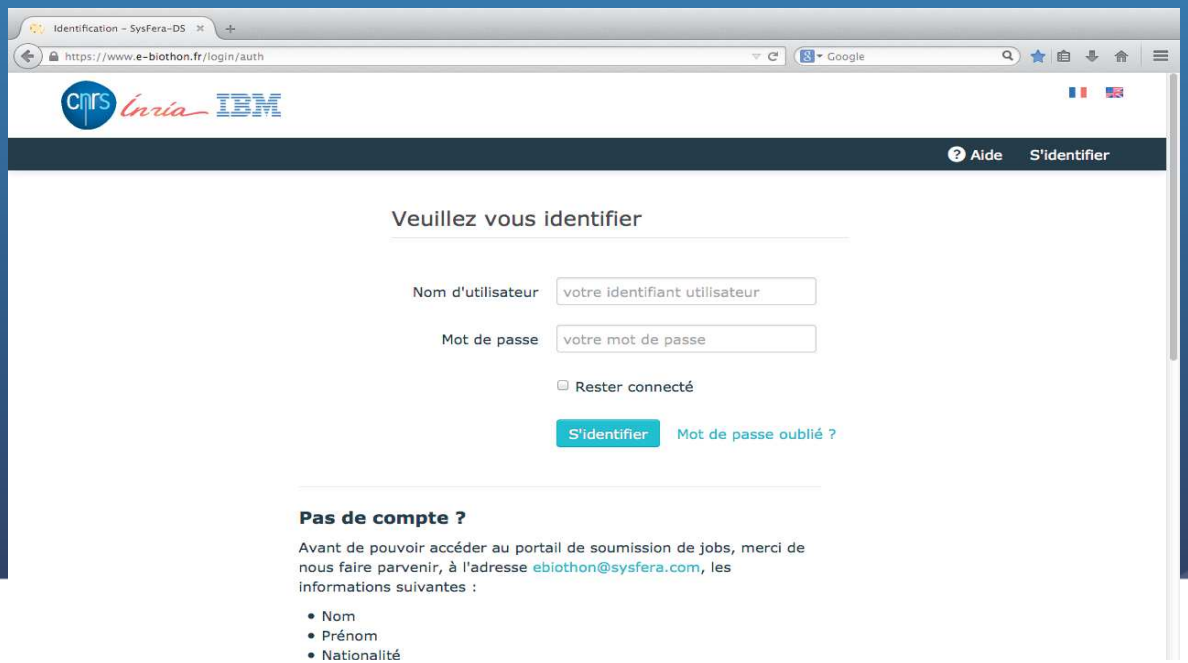


Projet PhyML

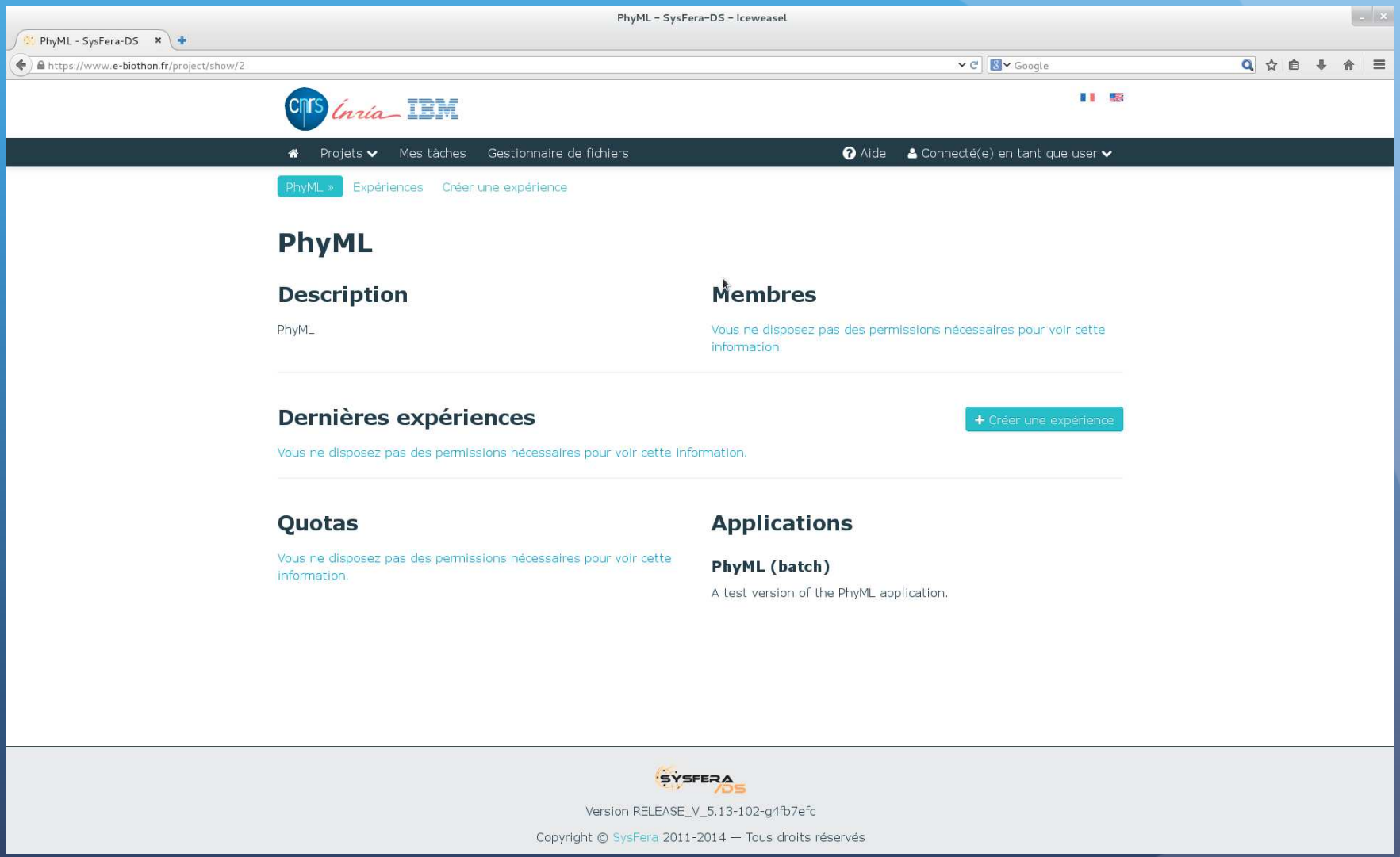
- Depuis octobre 2014, possibilité pour utilisateurs de PhyML d'employer la plate-forme E-Biothon lorsque leur demande ne peut pas être satisfaite par la plate-forme ATGC : http://www.atgc-montpellier.fr/phyml/
- Soumission de travaux via le portail SysFera-DS, à travers une interface dédiée réalisée 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

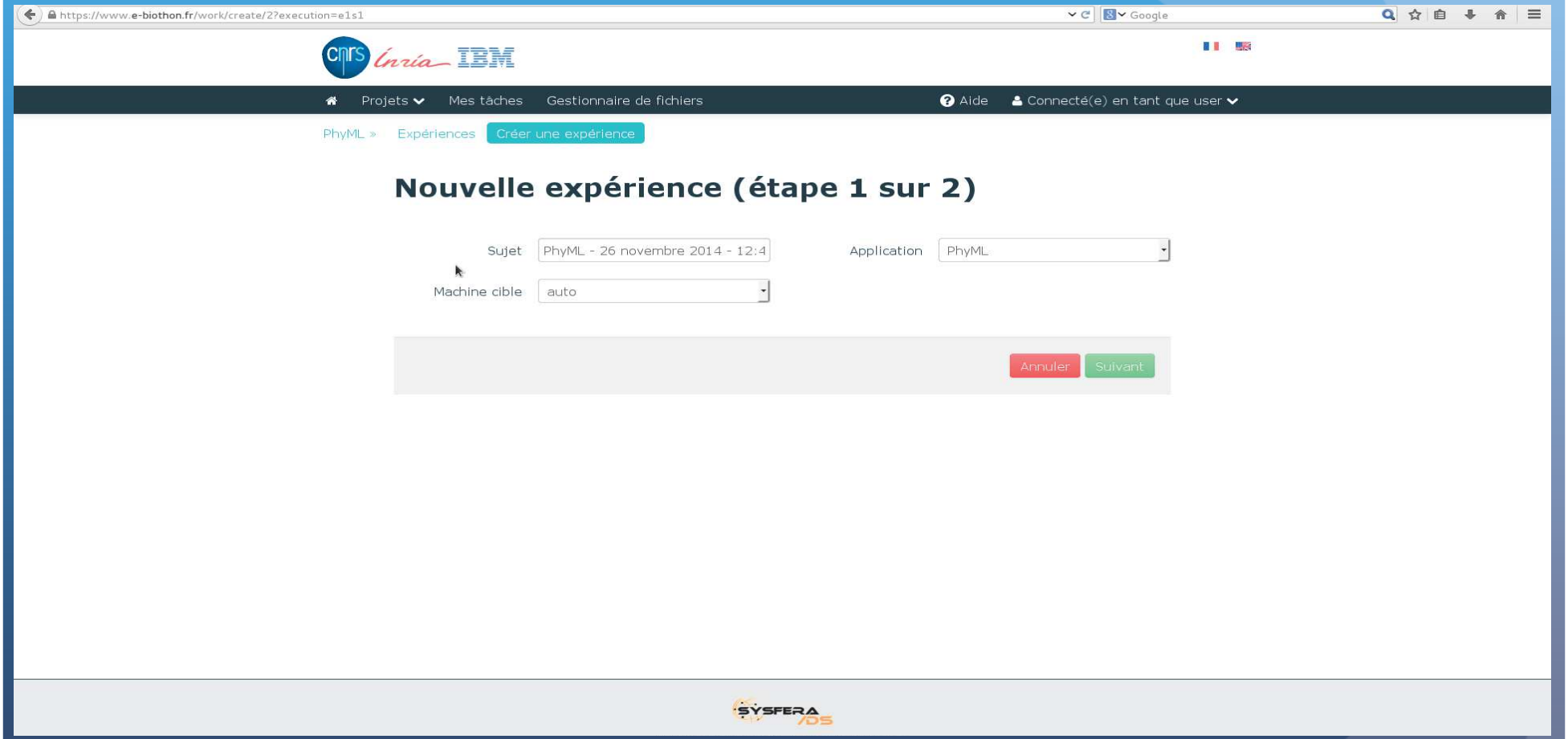
- Portail Sysfera-DS accessible à l'url suivante: https://www.e-biothon.fr/login/auth
- 1ère Connexion avec login et mot de passe provisoire pour le nouvel utilisateur.
- Personnalisation du mot de passe.



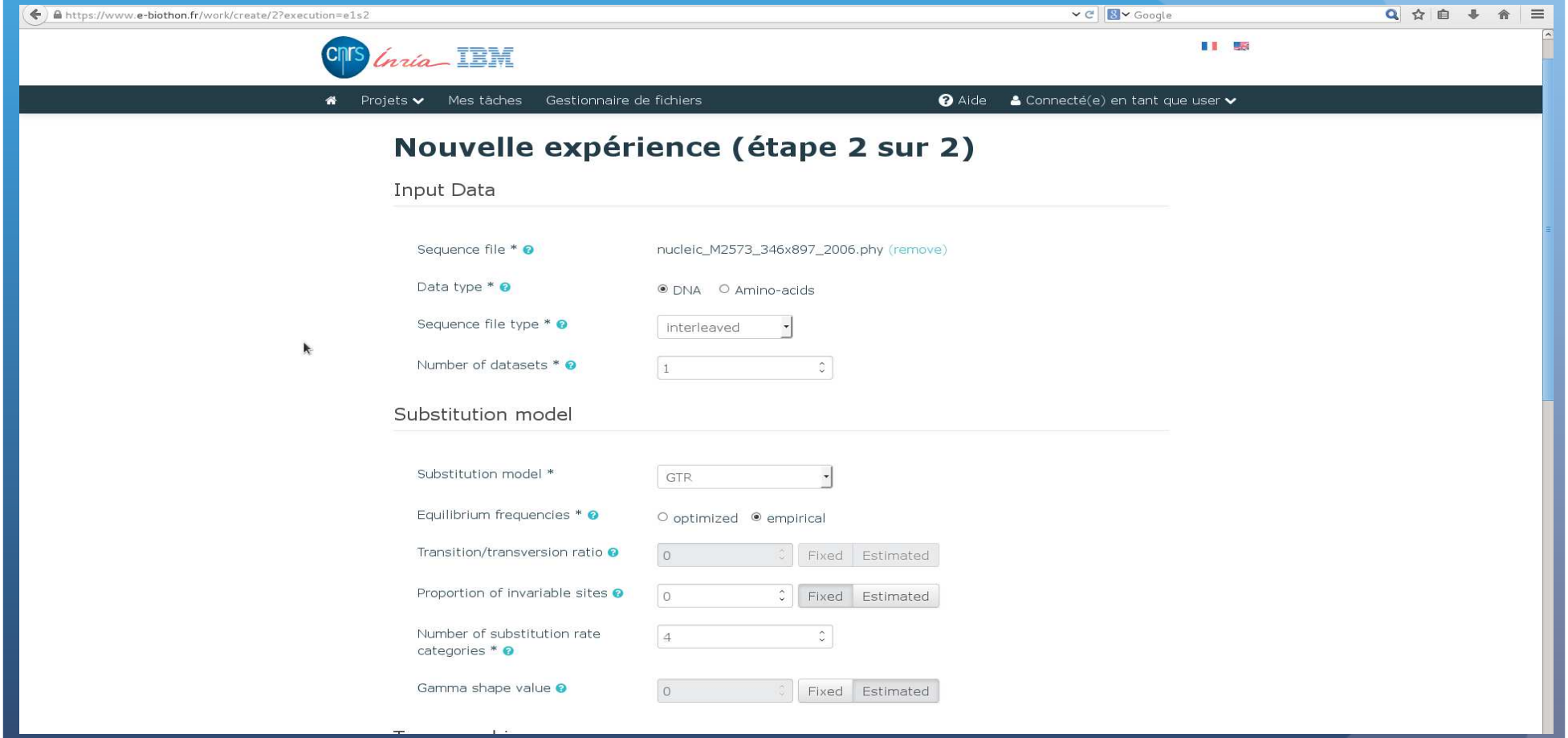
Page Accueil



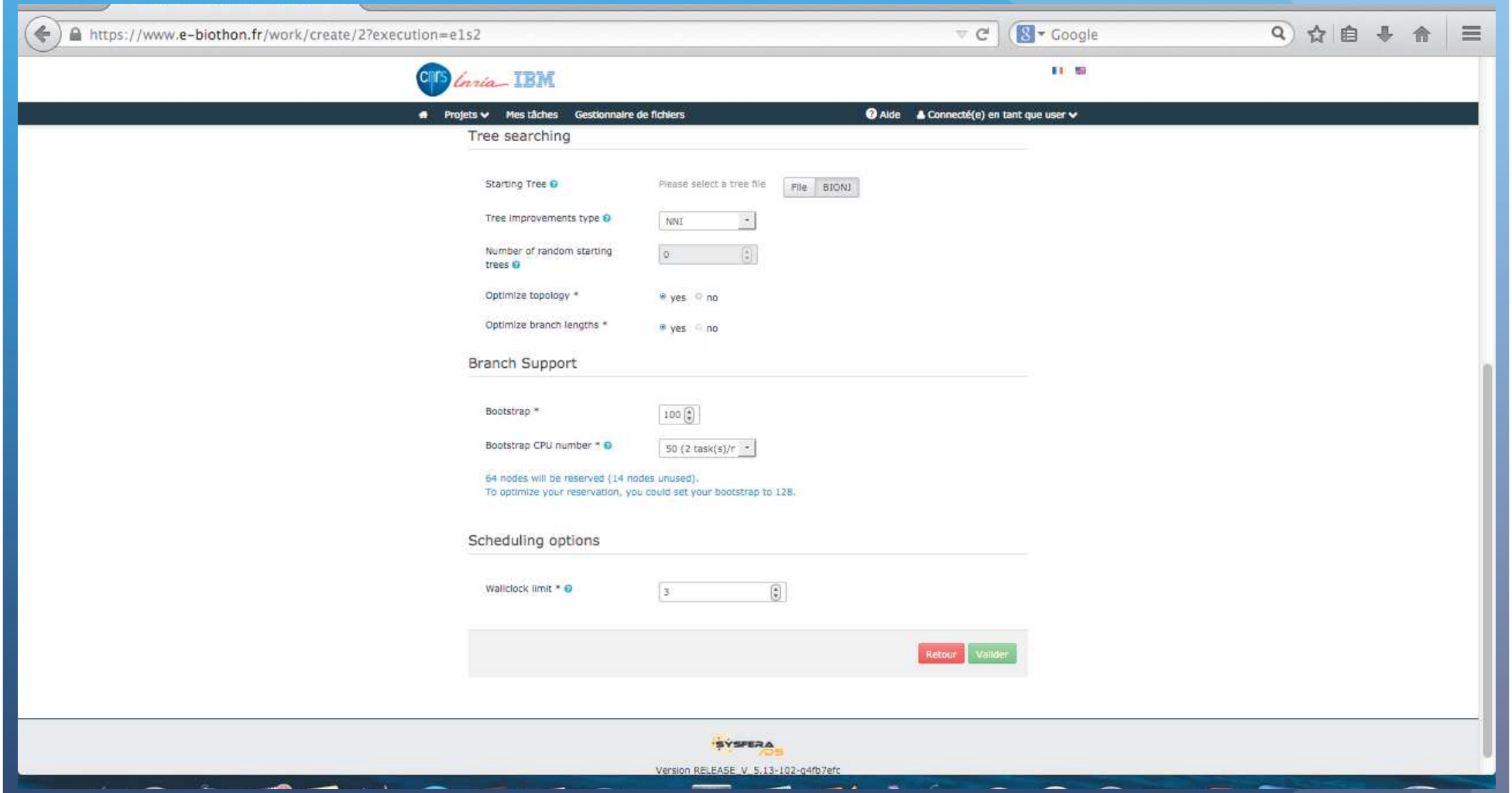
Créer une expérience 1/3



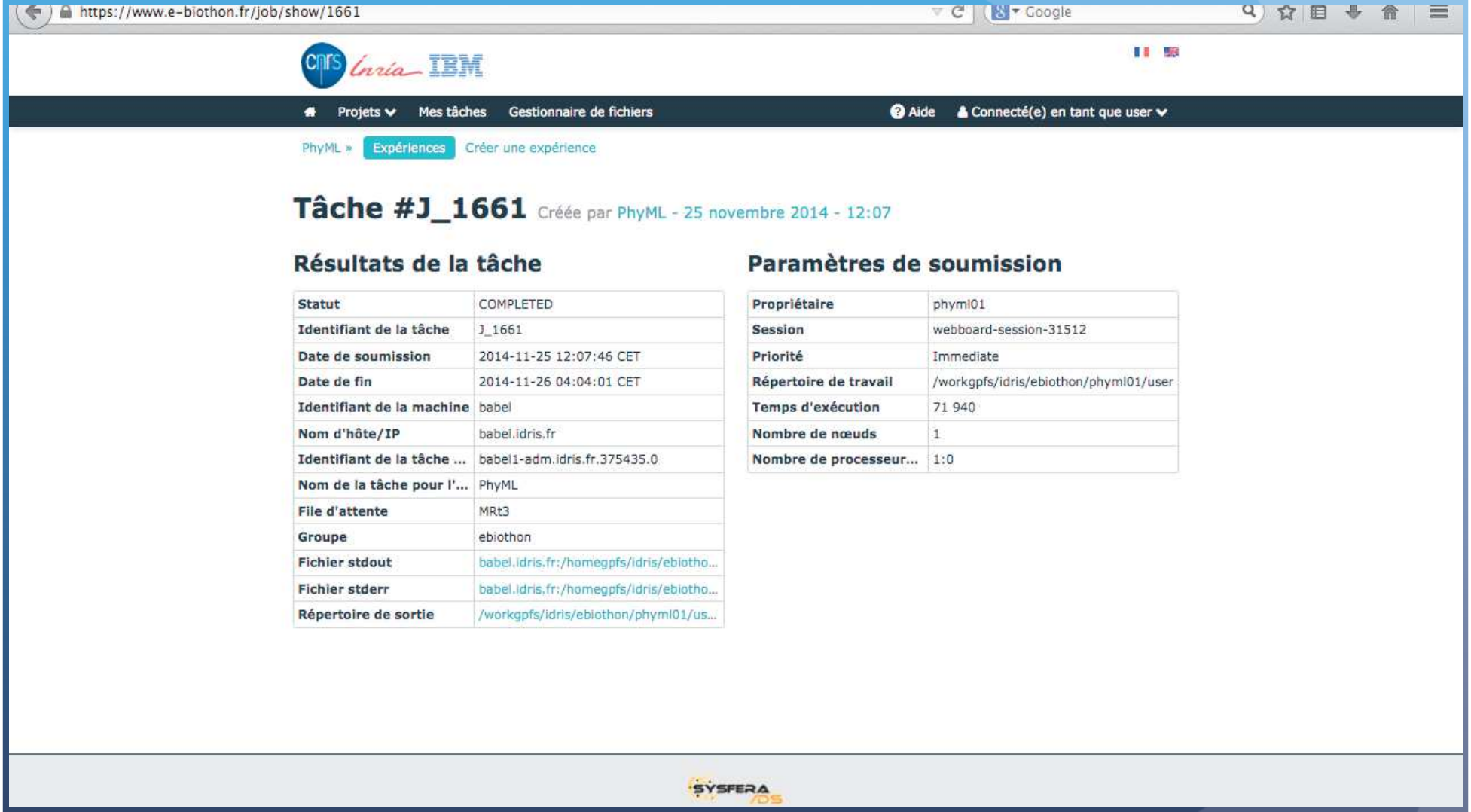
Créer une expérience 2/3



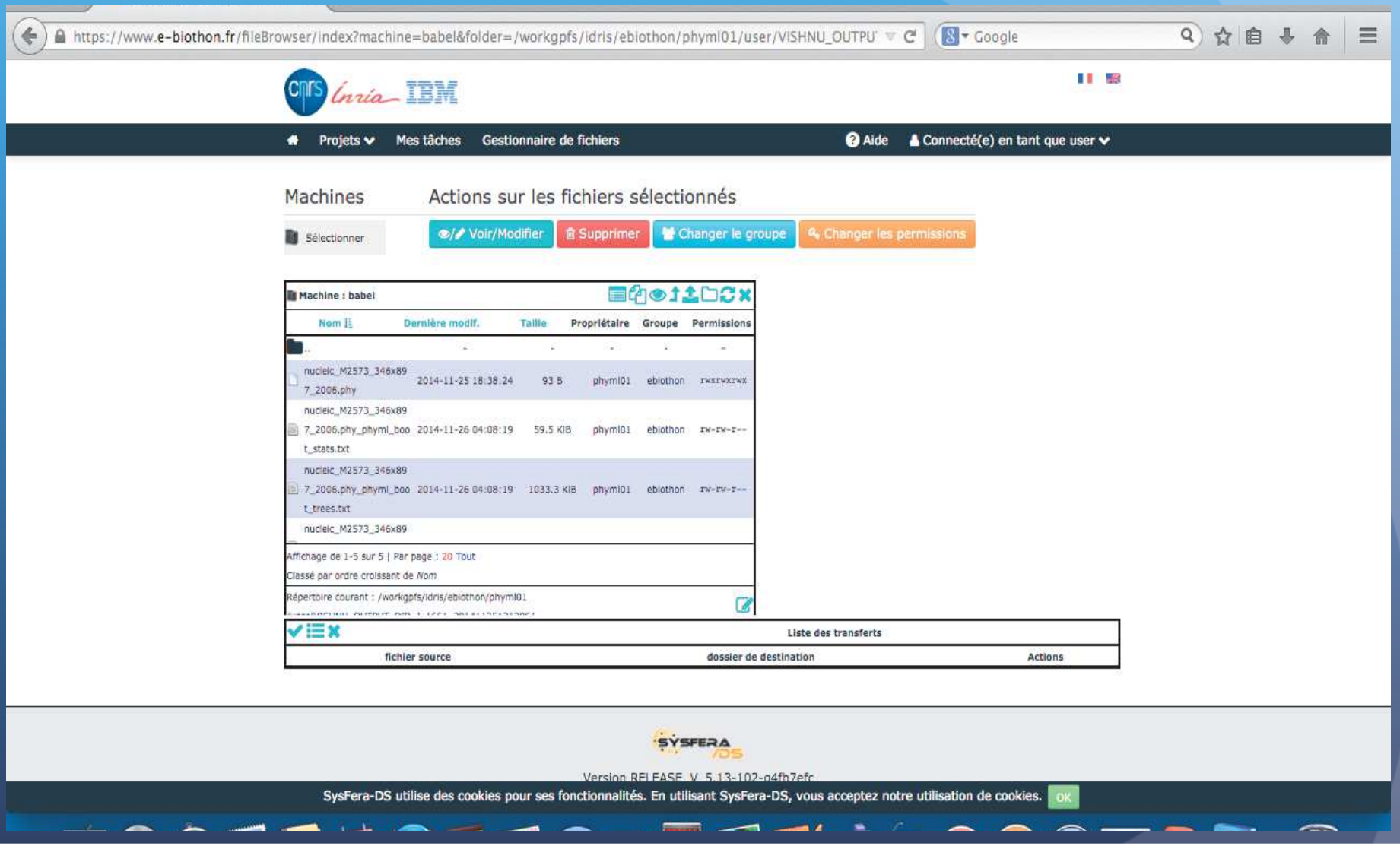
Créer une expérience 3/3



Suivi du statut



Résultats


https://www.e-biothon.fr/fileBrowser/index?machine=babel&folder=/workgpfs/idris/ebiothon/phyml01/user/VISHNU_OUTPU
Google
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 ebiothon rwxrwxrwx
nucleic_M2573_346x89 7_2006.phy_phyml_boot_stats.txt 2014-11-26 04:08:19 59.5 KiB phyml01 ebiothon rw-rw-r--
nucleic_M2573_346x89 7_2006.phy_phyml_boot_trees.txt 2014-11-26 04:08:19 1033.3 KiB phyml01 ebiothon rw-rw-r--
nucleic_M2573_346x89
Affichage de 1-5 sur 5 | Par page : 20 Tout
Classé par ordre croissant de Nom
Répertoire courant : /workgpfs/idris/ebiothon/phyml01
Liste des transferts
fichier source
dossier de destination
Actions
Version RELEASE_V_5.13-102-g4fb7efc
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,

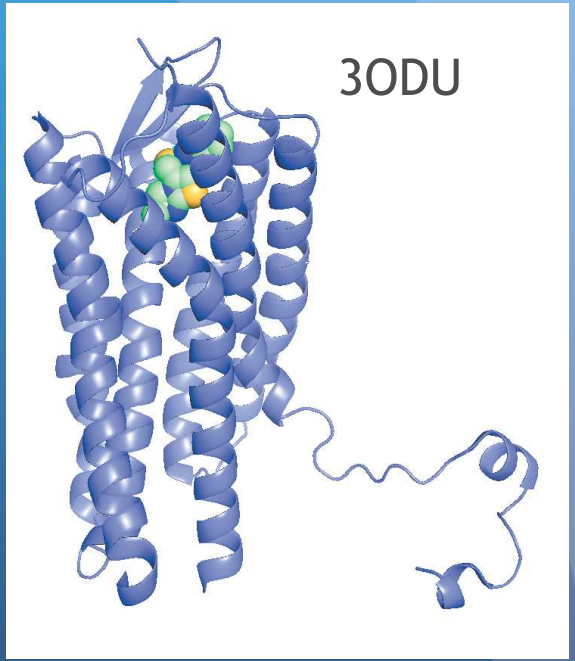
B. Taddese and M. Chabbert,
UMR CNRS 6214 – INSERM 1083, University of Angers, France

- 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

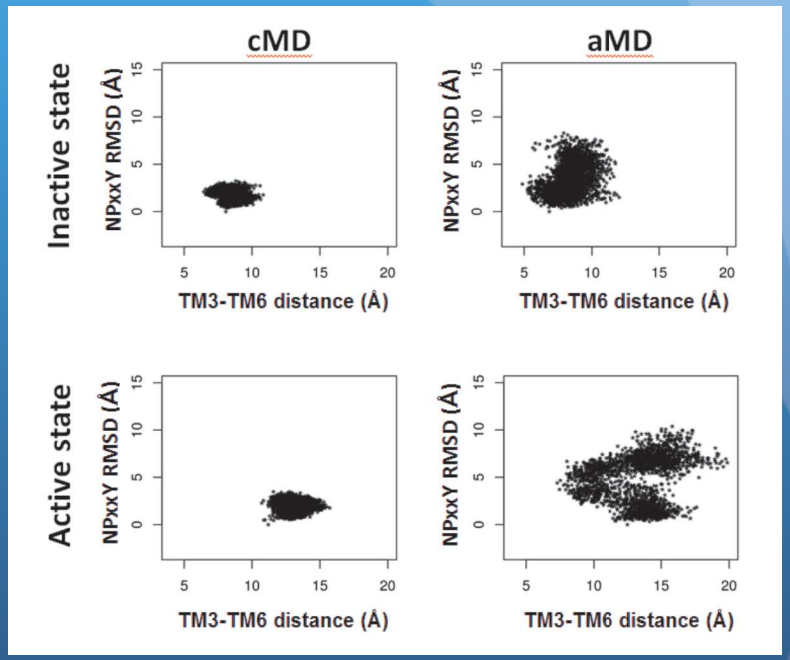
Prototype of chemotaxic receptors



Instituts thématiques Inserm Institut national de la santé et de la recherche médicale
cnrs
Agence Nationale de la Recherche ANR
NEC
université angers

- 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

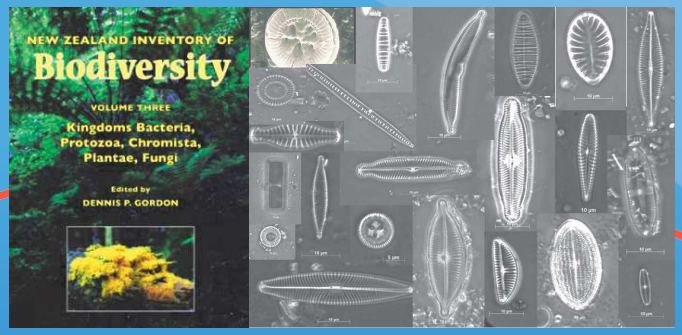
Conformational ensembles sampled by CXCR4 during a 60 ns trajectory obtained by classical (cMD) and accelerated (aMD) MD simulations at 310 K with NAMD.

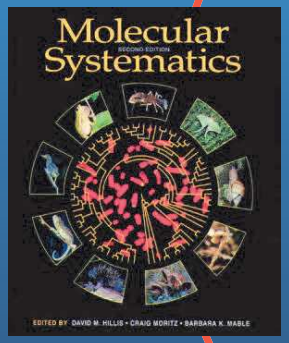


- System with ~63,000 atoms for monomeric receptor, POPC lipids, ions and TIP3 water molecules.
- Using E-Biothon with 512 CPU jobs, each nanosecond of

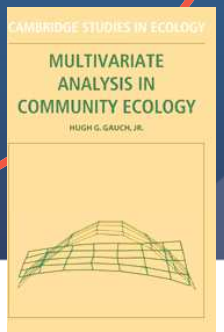
Biodiversiton project

Alain Franc
INRA - UMR BioGeCo /
Inria - Pleiade Team



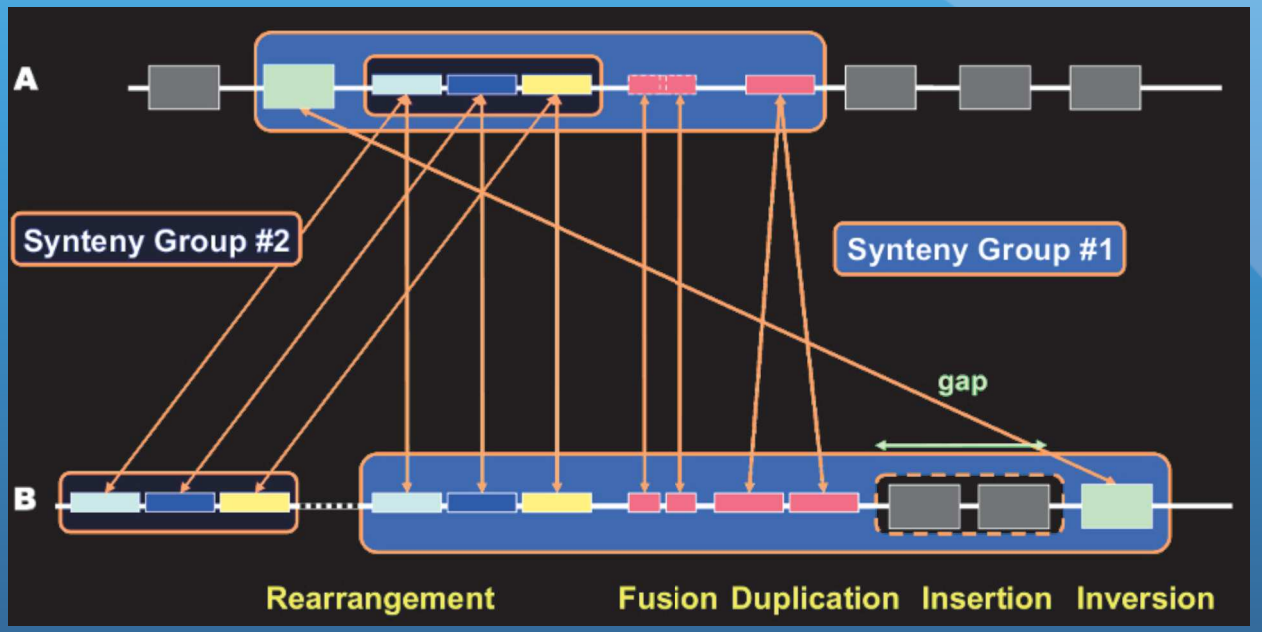


- Utilisation du parallélisme massif pour étudier la diversité biologique et la structure des différentes communautés d'organismes
- Parallélisation des calculs de distance avec MPI et passage sur le DARI (centres nationaux HPC)



Insyght: outil de comparaison des génomes procaryotes

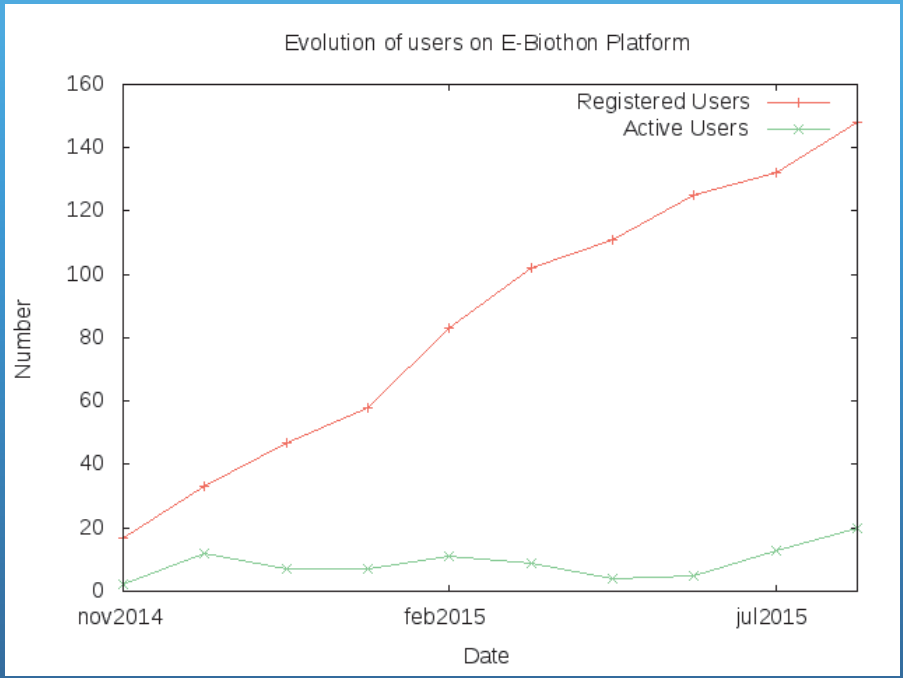
Jean-François Gibrat, INRA, UR1404, Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement

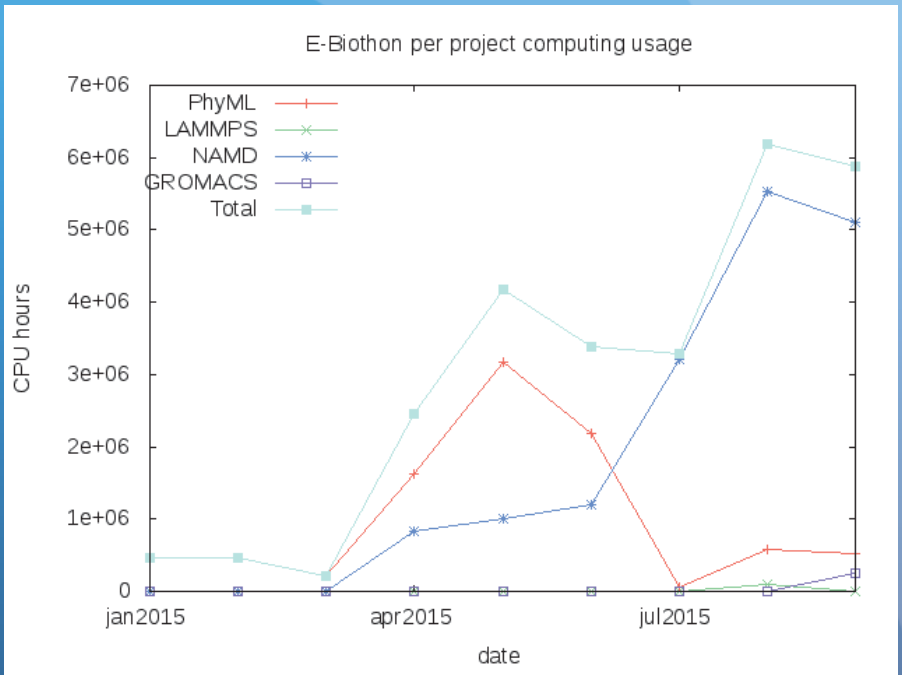


- Comparaison du contenu en gènes de près de 2700 génomes complets de bactéries avec environ 2 mois de la plateforme complète (>3.5 M comparaisons de génomes chacun avec 16 M de gènes, BD de 3.5 To)
- Permet de visualiser les séquences protéiques homologues et les relations de synténie chez les procaryotes 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, dextraction des relations de synténie et visualisation facilement grâce à une interface web accessible à tous.

Utilisation de la plateforme

Utilisateurs et charge

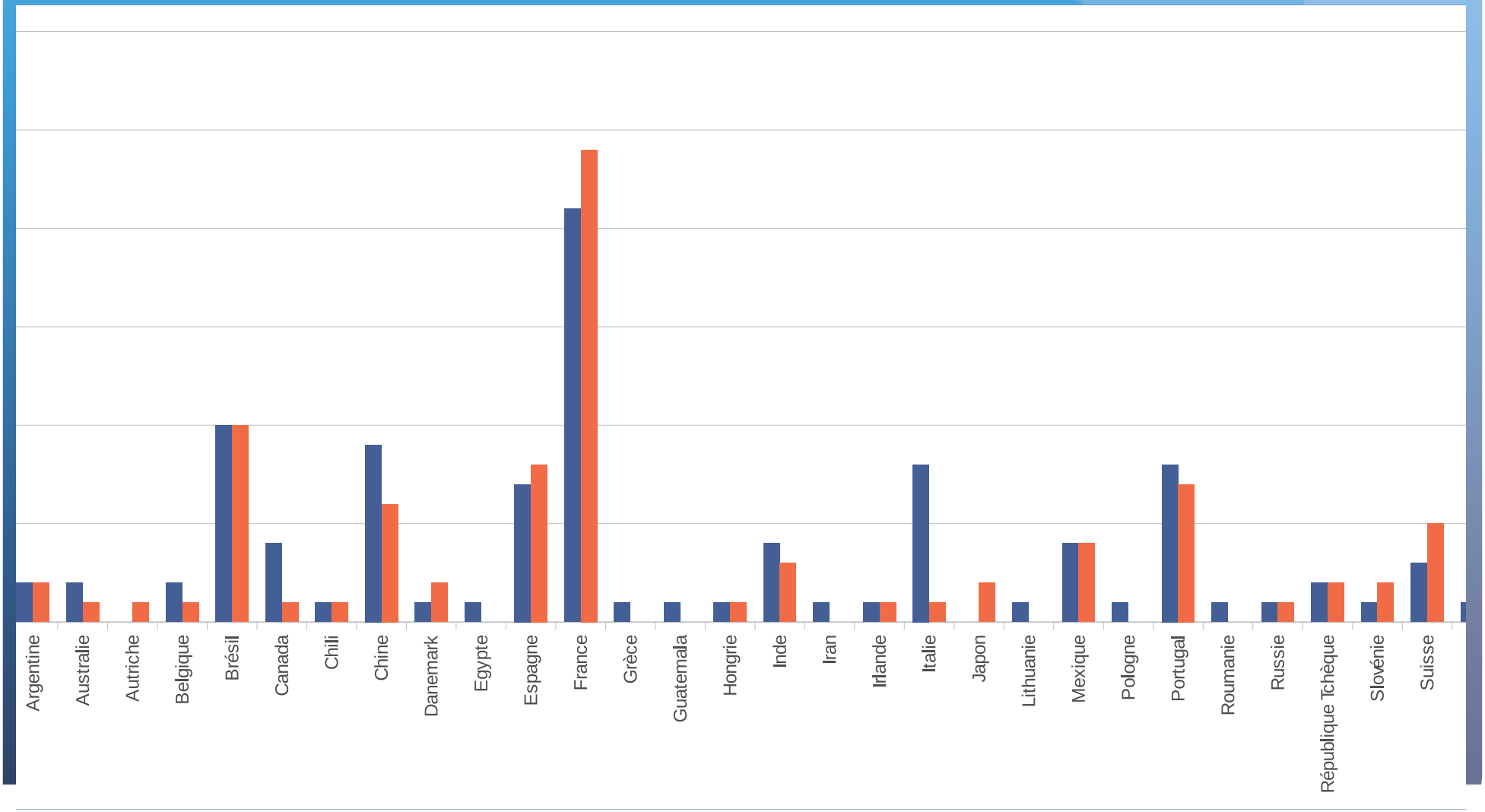




Mai 2016 :

- 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 %)

Provenance des utilisateurs



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 opérationelle
- 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 Francais de Bioinformatique and SysFera)
  - Crucial pour déployer de nouvelles applications et gérer la plateforme
- *Remerciements à CNRS, IBM, IDRIS, Inria, l'Institut Francais de Bioinformatique and SysFera sans qui ce projet n'aurait pas vu le jour et aux utilisateurs* !