E-Biothon : Une plate-forme pour accélérer les recherches en biologie, santé et environnement.

N.Bard, S.Boin, F.Bothorel, P.Collinet, M.Daydé, B. Depardon, F. Desprez, M.Flé, A.Franc, J.-F. Gibrat, D. Girou, P.-F. Lavallée, V.Lefort, M.Rugeri, E.Ruinet, C.Séguin, S.Thérond.

Journées SUCCES – 13 Novembre 2013

Introduction

- **Des constats**
  - Des besoins grandissants autour des sciences du vivant (parallélisation, gestion de données, algorithmique adaptée, …)
  - Un besoin de transparence dans l'utilisation des ressources informatiques
  - Des plates-formes disponibles et des logiciels matures pour les gérer
  - Le mode SaaS (Software as a Service) s'impose maintenant même pour les applications scientifiques

Bioinformatique : des besoins croissants

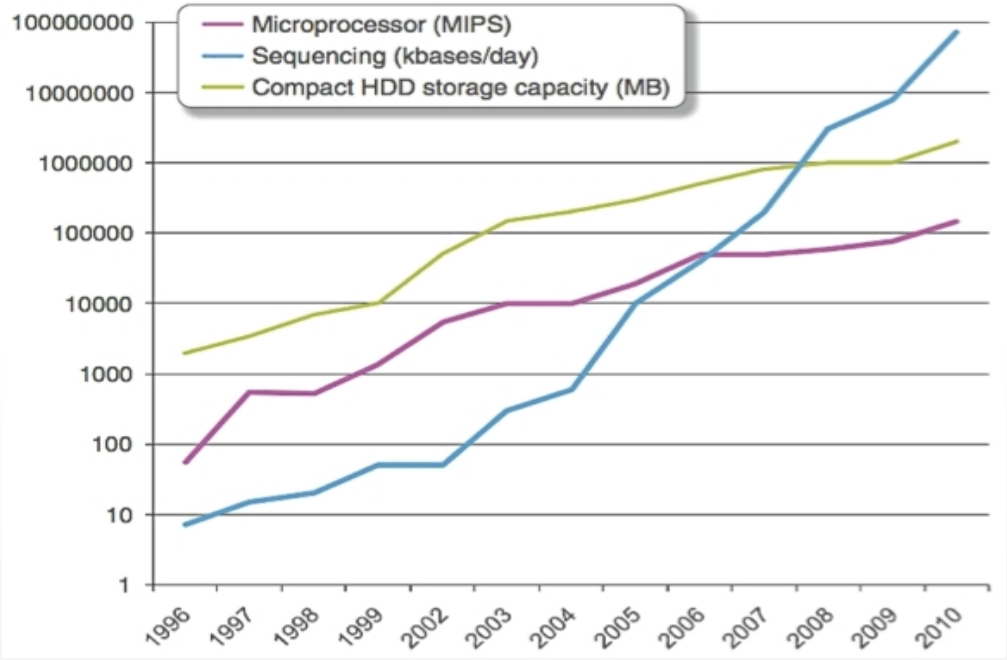
Augmentation du nombre et de la taille des données

- De plus en plus de données séquencées
- Coût des séquenceurs qui décroit et leur nombre qui croit

Impact de la modélisation dans la biologie

- Nombreuses équipes qui travaillent sur des modèles mathématiques et sur des programmes informatiques qui modélisent les phénomènes et analysent les données

```
>inconnu
GGTACATCCGCCCTTATTTAAATTTCGAAGATTTACCATT
GAATTTTACCTTTCGAATTTGACCTTTACTTAATAAGGTA
AAAATAAGATAAAAAAAAATTCGAAAGCTTTATCTTTTTT
TTCCCATCTAAATTTGTATTTCAATTTTAAAAATAACGTA
CCAATAATCCTCAATAATAACCCAACCCATAAGAATTTGA
TTATTTTGTTGAAATAAAAAAAAACTATATTTAAAGAAAT
CGAGAAAAACTCATTTACGTCAAGTAGTAAAGTAAAAAAA
TACTAAAAAAAAAGGAGCAATAGCCAAGTCAATTTCCTGTT
CTAGGGAGAAGATTGGTATTGCTCCTTTTTTTTTTCAAAA
ACTTGTATCTATTATCTATTAAATTAAAACCGAAACCTTA
TCCAGAAGTTTTATCCATTTGTAGATGGAGC
```



Partenaires

Rôles des partenaires

- **CNRS**
  - pilotage du programme scientifique
  - participation de ses équipes UMR ou UPR à la réalisation de recherches dans le projet
  - participation à la mise en place, le bon fonctionnement des moyens de calcul ainsi que l'aide au déploiement des applications et le support utilisateur.
- **IBM**
  - apport technologique, la mise en place d'éléments d'une plateforme distribuée
  - apport scientifique par la collaboration avec ses laboratoires de recherche.
- **INRIA**
  - participation de ses équipes à la réalisation de recherches du projet
  - aide au déploiement des applications et le support utilisateur
- **IFB**
  - rôle « d'intermédiaire » entre la communauté des sciences de la vie et la communauté de la recherche en informatique et bioinformatique pour contribuer au relais de cette action vers la communauté des sciences de la vie et l'identification d'applications pertinentes
- **SysFera**
  - fourniture de la solution logicielle de gestion de la plate-forme et du passage en mode SaaS

L' héritage du Décrypthon

Objectifs

- Accélérer la recherche sur les maladies génétiques et rares.
- Rendre transparente l'utilisation de ressources de calcul distribuées aux utilisateurs.

Moyens

- Une grille de calcul comportant six sites (Bordeaux 1, Lille 1, l'ENS de Lyon, Paris IV, Orsay, et l'UPMC)
- L'intergiciel DIET, et une interface web.

BlueGene/P : le cœur de la plate forme

Deux racks de Blue Gene P.

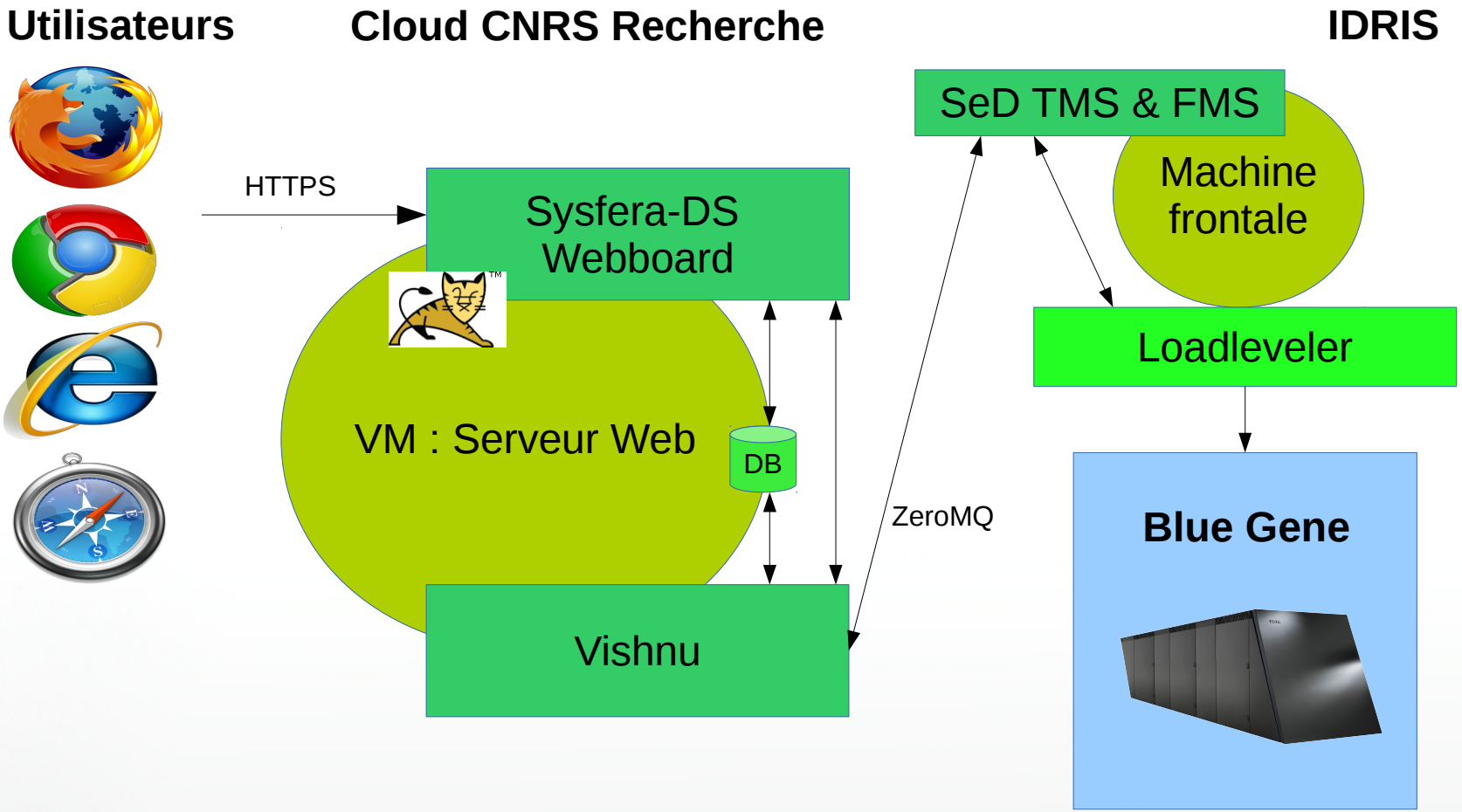
- Puissance en crête de 28 téraflops
- Chaque rack compte 1024 nœuds de 4 cœur.
- Chaque nœud possède 2 gigaoctets de mémoire RAM.
- Une capacité de stockage de 200 téraoctets.

Deux modes de fonctionnement

- Le mode standard.
- Le mode High Throughtput Computing.

La solution Sysfera

Telle qu'elle est déployée sur notre plate-forme.



SeD : server daemon, un processus de service de vishnu.

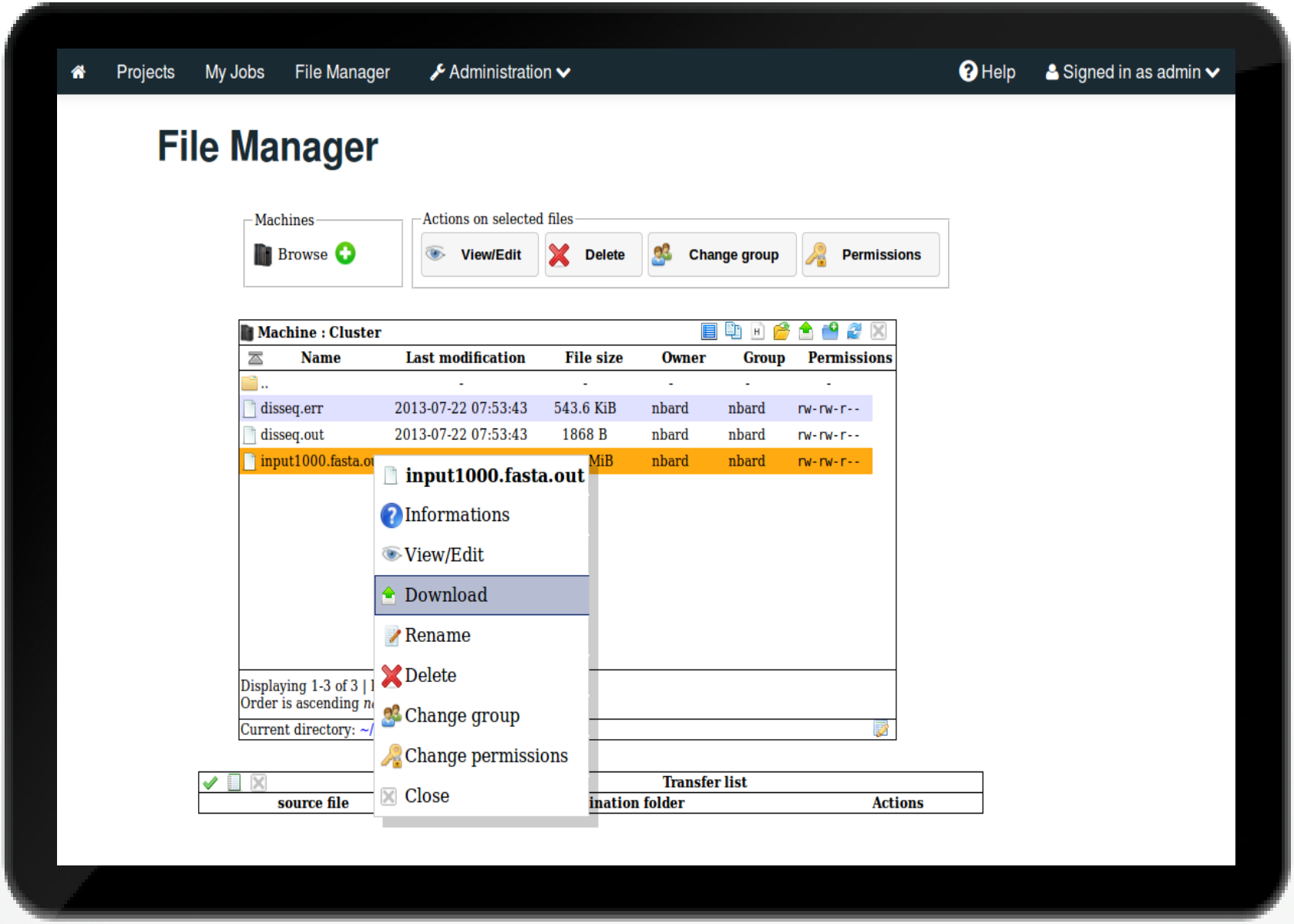
WebBoard : l'interface Web

Gestion des utilisateurs
- Authentification des utilisateurs.
- Gestion en projets.

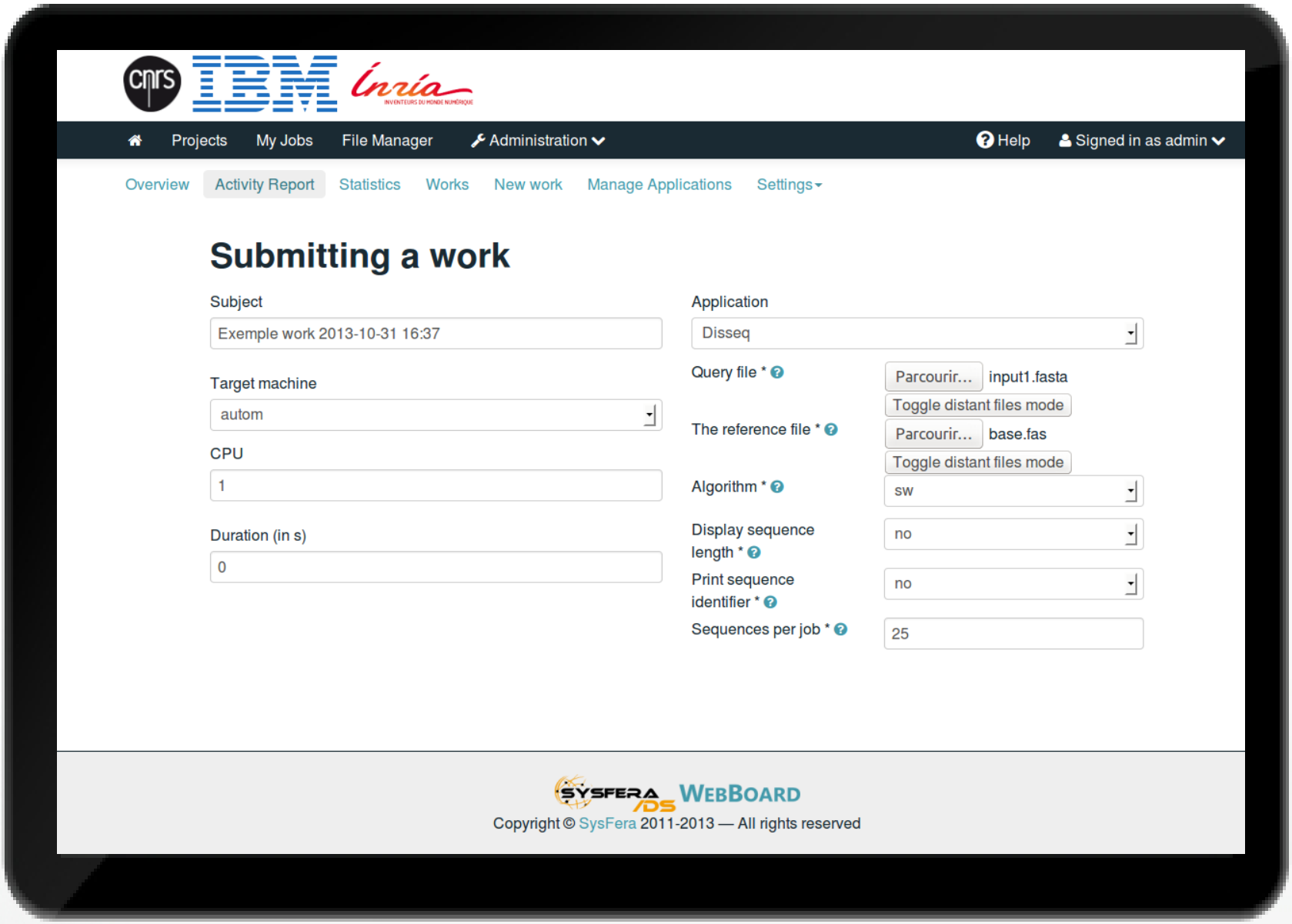
Gestion des droits
- Un système de rôles auxquels sont associées des permissions.
- Des rapports d'activités pour les administrateurs.

Gestion des applications
- Les applications sont facilement entrées dans la base de données.
- Les applications peuvent être liées à des projets.
- Les applications peuvent être mises à jour « en temps réel ».

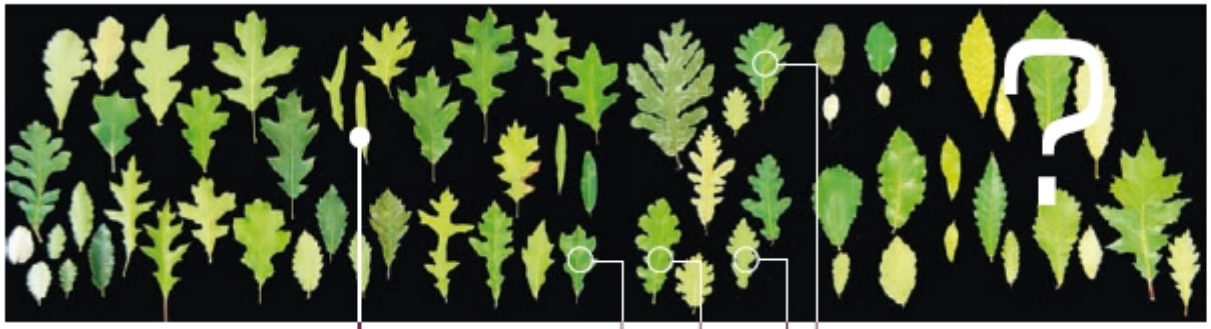
Gestion de fichiers



Soumission de jobs



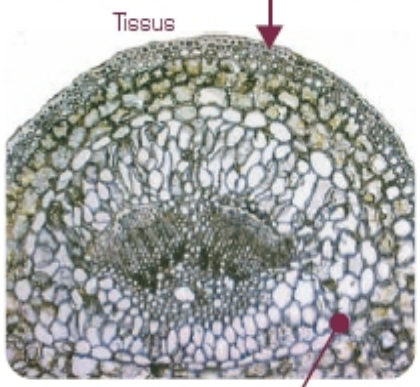
Disseq

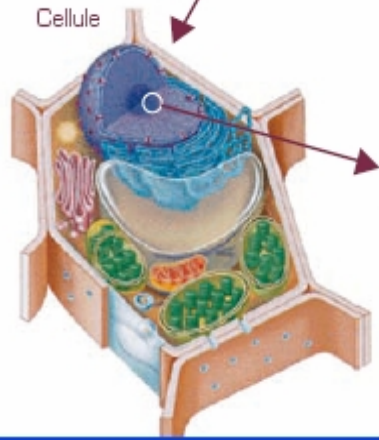


La diversité des phénotypes (ex : forme des feuilles) pose des problèmes d'identification des différentes espèces

La diversité des génotypes (différences de séquence) peut différencier les espèces de chênes

ACGTGTGCTAT ▶ *Quercus petraea*
ACGCGTGCTAT ▶ *Quercus robur*
ACGT--GCTAT ▶ *Quercus pubescens*
ACGCAGTCTAT ▶ *Quercus cocinea*





Extraction,
Amplification,
Séquençage
ADN

Les séquences d'ADN permettent de mettre un nom sur un organisme, de la bactérie à l'homme comme les empreintes digitales ou l'iris de l'œil permettent de reconnaître un individu.

Certaines séquences sont propres à chaque individu, d'autres sont partagées par un groupe d'individus.
Le défi consiste à identifier ces séquences partagées par tous les individus d'une même espèce, d'un même genre ou d'une même famille.

Taxonomie et distance d'édition

Définition: La **distance d'édition** entre deux chaînes de caractères est définie comme :

Le nombre minimal de modifications requis pour passer d'une chaîne à l'autre, les opération de modification possibles étant l'insertion, la suppression, ou la substitution d'un caractère.

kitten → **s**itten (substitution de 'k' et 's')
sitt**e**n → sitt**i**n (substitution de 'e' et 'i')
sittin → sittin**g** (ajout de 'g' à la fin).

SOVIET PHYSICS–DOKLADY

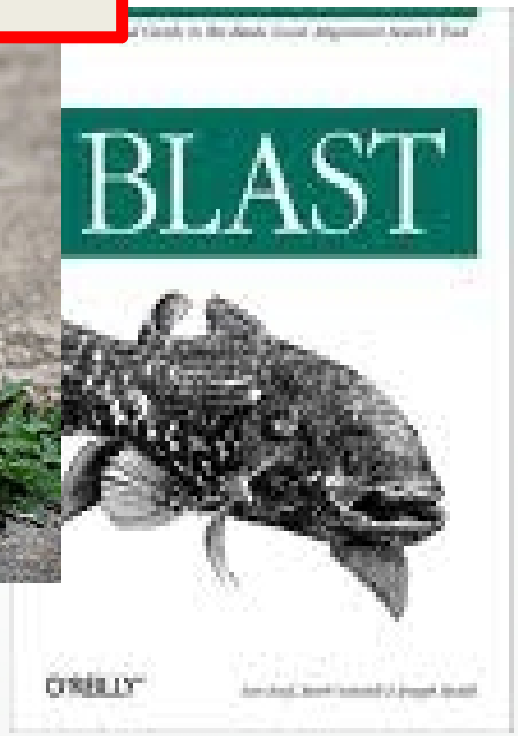
BINARY CODES CAPABL
DELETIONS, INSERTIONS

V. I. Levenshtein

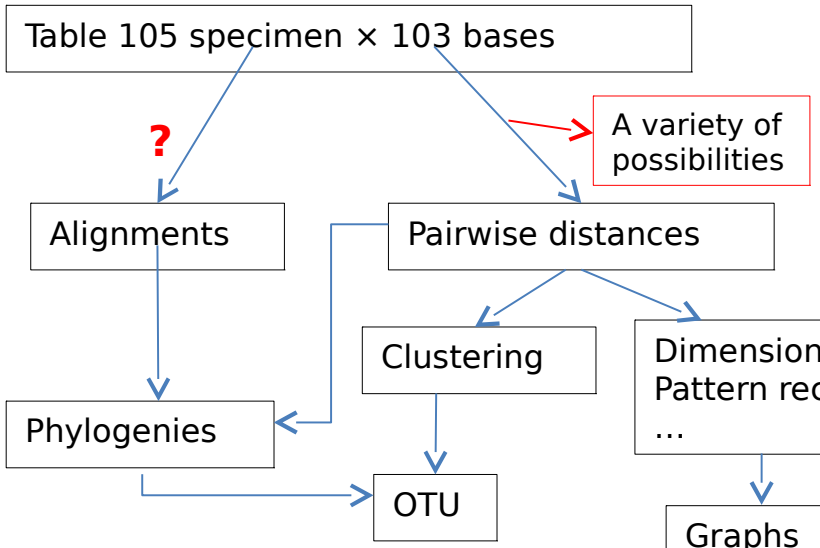
(Presented by Academici
Translated from Doklady
pp. 845-848, August, 196
Original article submitte

Investigations of transmission of binary information usually consider a channel model in which failures of the type 0 → 1 and 1 → 0 (which we will call reversals) are admitted. In the present paper

were inserted (deleted) from at least one of the words x or y to obtain z are deleted from (inserted into) the word z, then, as we can easily see, we obtain a word that can be obtained from both x and y



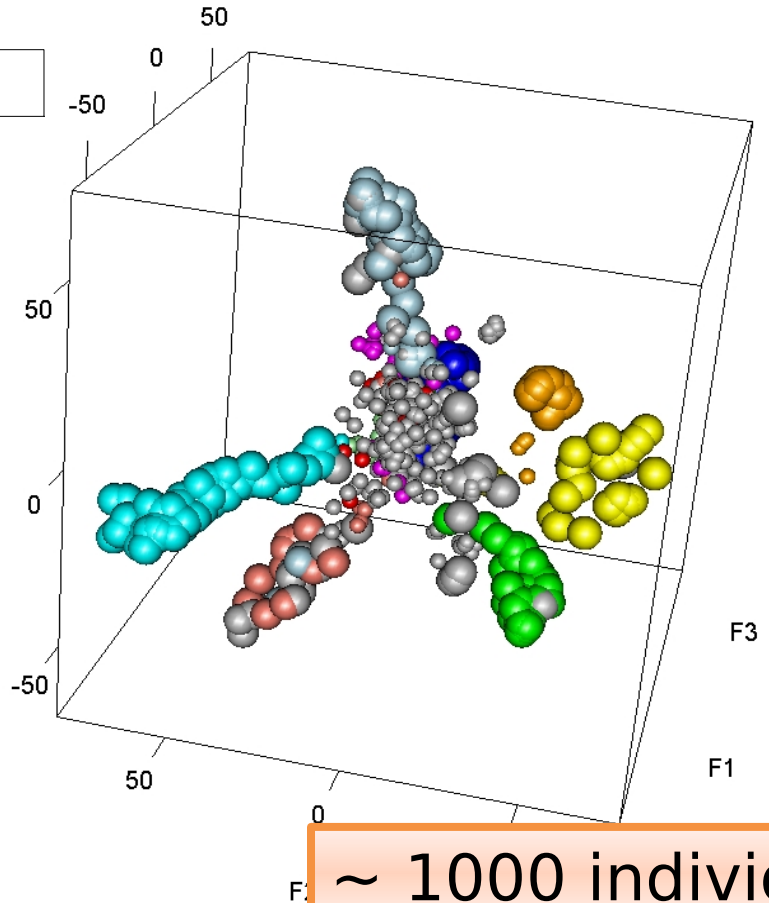
Application



La théorie (botanique) indique une structure hiérarchique de la diversité espèces – genres – familles – ordres …

La retrouve-t-on dans les distances entre séquences ?

blue -> Mimosoideae
---
lightblue -> Lecythidaceae
---
cyan -> Chrysobalanaceae
---
green -> Annonaceae
---
lightgreen -> Caesalpinioideae
---
yellow -> Myrtaceae
---
orange -> Elaeocarpaceae
---
magenta -> Apocynaceae
---
salmon -> Burseraceae
---
red -> Malvaceae

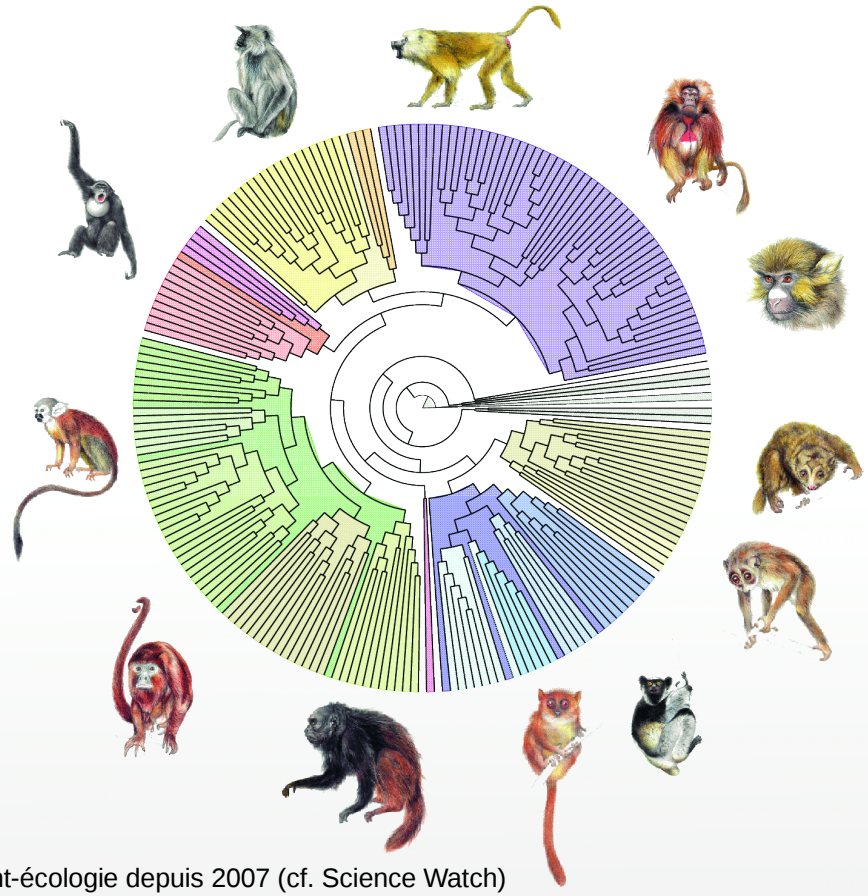


~ 1000 individus

PhyML

- Un outil de comparaison des espèces du vivant, efficace, rapide et précis.
- Basé sur le principe du maximum de vraisemblance.
- L'algorithme calcule la probabilité que les données correspondent à plusieurs modèles d'arbres d'évolution darwinienne, et le résultat est le modèle ayant la plus haute probabilité.

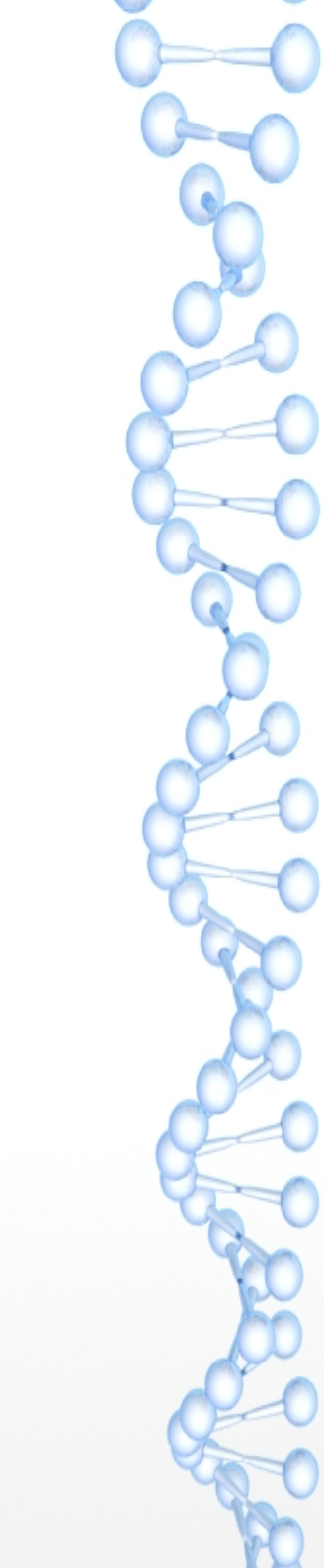
Original algorithm :
(a) A first tree is built by a fast distance based algorithm (BIONJ).
(b) The model parameters (e.g., gamma) are optimised and periodically updated.
(c) The tree is iteratively refined until convergence:
(1) compute all possible changes;
(2) apply a proportion λ of these changes;
(3) check that the modified tree is better than the current tree, otherwise divide λ by 2 and return to (2).
(d) Return the current tree.



La publication de PhyML est la plus citée au monde en environnement-écologie depuis 2007 (cf. Science Watch)

Perspectives

- Mise en place de la plate-forme et de son environnement logiciel
- Prise en compte des particularités de la plate-forme matérielle
- Validation sur 3 premières applications
- Appel à projet en prévision pour 2014 avec un ouverture vers la communauté sciences de la vie
- Possibilité d'ajouter d'autres ressources à la plate-forme
- **Contacts**
  - Michel Daydé (CNRS) : michel.dayde@irit.fr
  - Frédéric Desprez (INRIA) : Frederic.Desprez@inria.fr
  - IFB



Merci de votre attention.

Avez-vous des questions ?