Quelle place pour la bioinformatique à l'IN2P3 ?

V. Breton

Journées informatiques

Cargèse 07/01

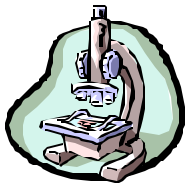
Définitions

- Bioinformatique : informatique d'acquisition et d'analyse des données issues de la génomique et la postgénomique
- Génomique : décodage du patrimone génétique de tous les êtres vivants (lecture de l'ADN)
- Post-génomique : ensemble des expériences biologiques pour comprendre la fonction des gènes
  - Transcriptome : décodage de l'ARN
  - Protéome : caractérisation des protéines
  - Métabolome : caractérisation du métabolisme

Genome sequencing projects
(sequencing=ATGCCCATGGTACCACCCTATGG...)

- Bacteria: 29 complete genomes (19 in the last 12 mounths)
- Archaea: 6 complete genomes
- Eukaryotes: 3 (4) complete genomes
  - yeast: 13 Mb 100%
  - *P. falciparum* 30 Mb 24%
  - *C. elegans* 100 Mb 95%
  - *A. thaliana* 120 Mb 60%
  - *Drosophila* 170 Mb 60% (100%)
  - human 3200 Mb 30%
    - « draft » complete in 2001, finished in 2003
  - mouse 3000 Mb 1%

Human Genome Project



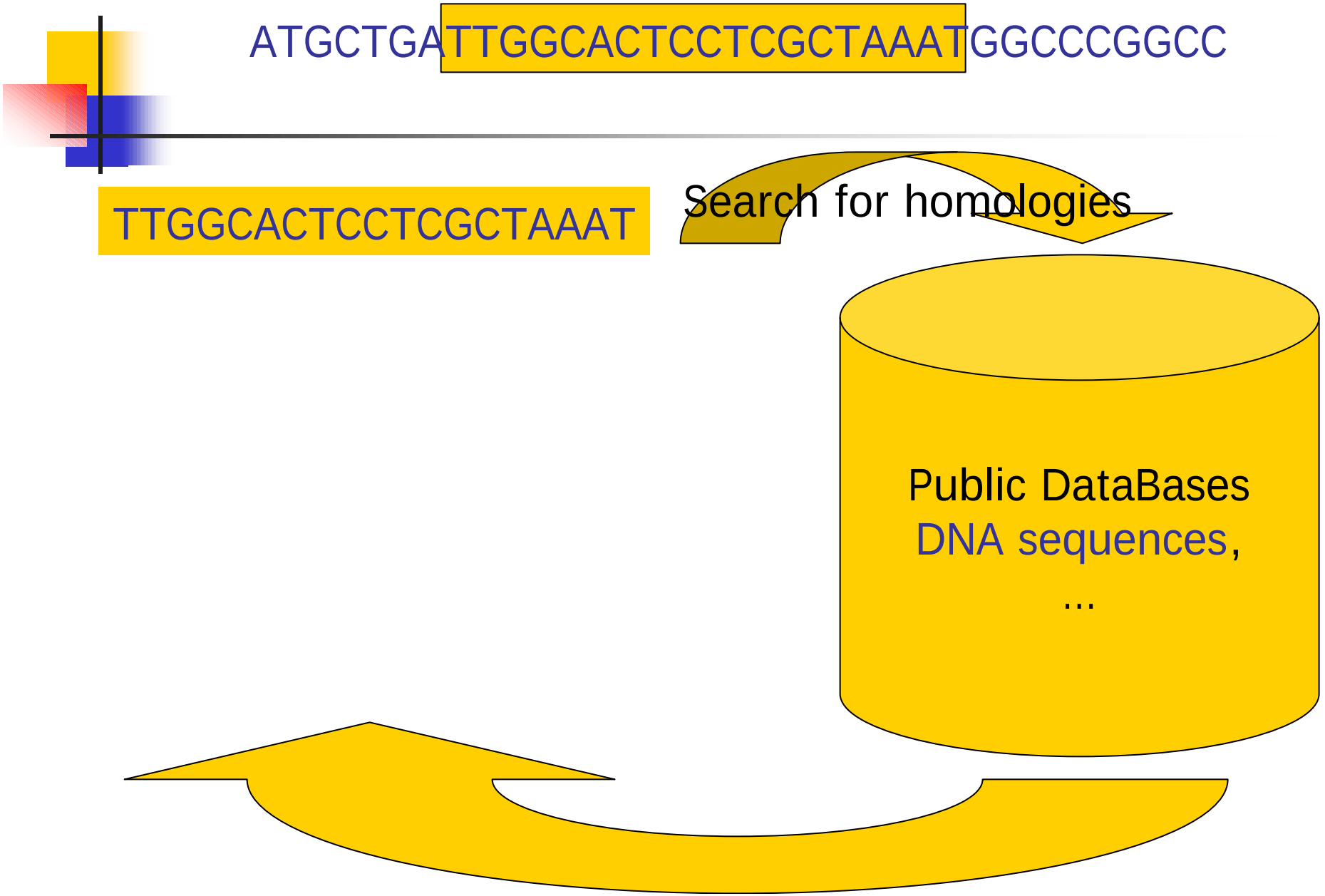
Séquençage

ATGCTGATTGGCACTCCTCGCTAAATGGCCCGGCC...

X B D Y U E F T H

Annotation

ATGCTGATTGGCACTCCTCGCTAAATGGCCCGGCC


ATGCTGATTGGCACTCCTCGCTAAATGGCCCGGCC
TTGGCACTCCTCGCTAAAT
Search for homologies
Public DataBases
DNA sequences,
...

Next steps in genome projects

- Identify genes and other functional elements within genomic sequence (where are the genes ?)
- Determine the function of genes (what do they do ?)
- Analogy with particle physics in the 60's and 70's
  - Plenty of data, new particles
  - Need for a strong theoretical foundation

Where are the genes ? What is their function ?

Experimental characterization of all human genes.

- 30,000-100,000 genes: How many years ?

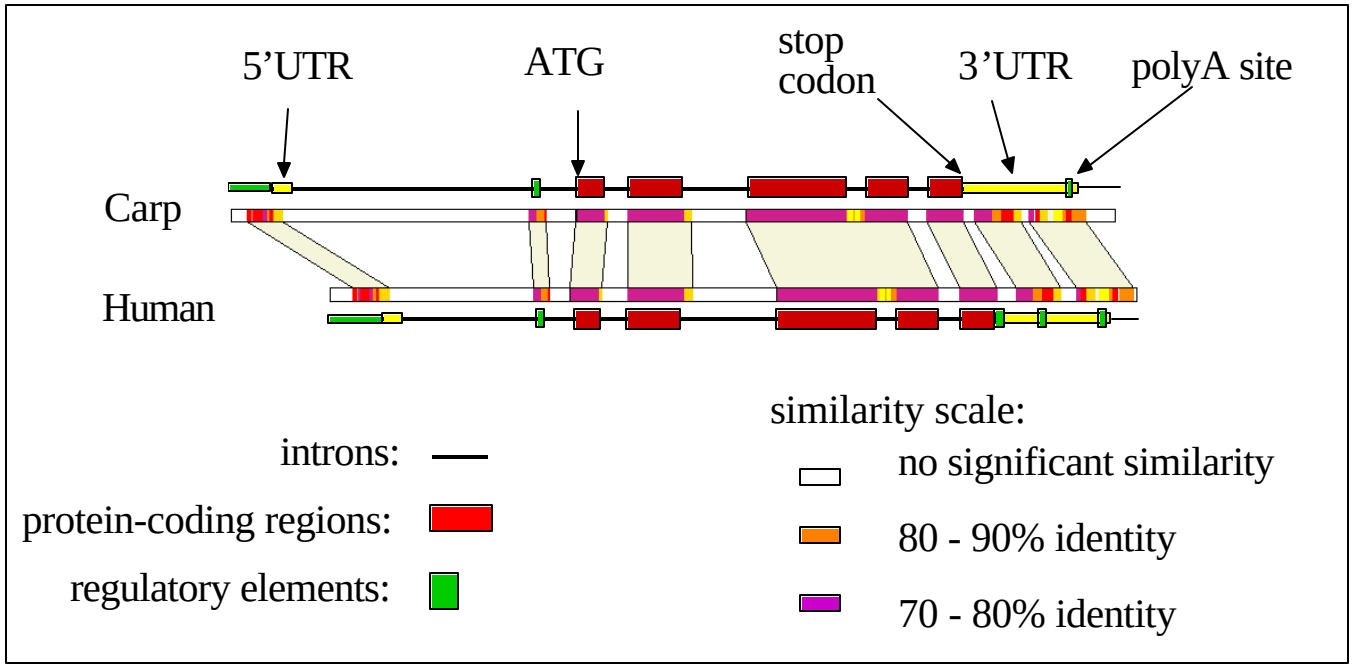
Computer analysis of genomic sequences

- (Ab initio methods)
  - Ruled-based or statistical methods
  - e.g.: coding sequence prediction, promoter prediction, …
  - Very useful but … limits in sensibility/specificity

- Comparative sequence analysis

Comparative analysis of homologous (= evolutionary related) sequences

Identification of functional elements in genomic sequences



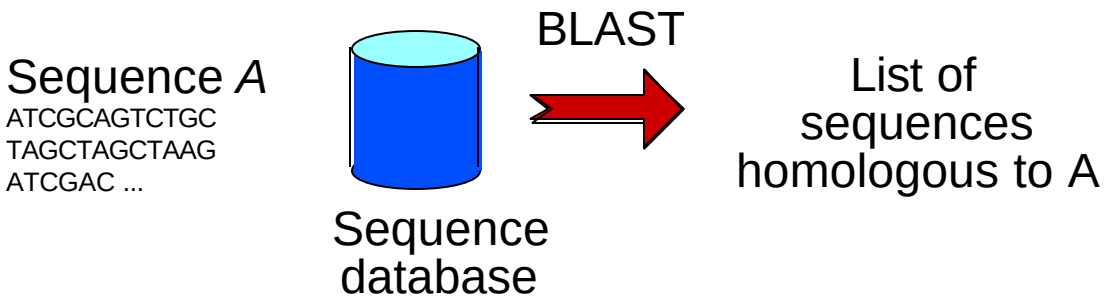
Prediction of gene function

Gene *A*: known function (*F*)

Gene *B*: homologous to *A* => function similar to *F*

BioInformatic Tools for Comparative Sequence Analysis

- (1) Similarity search programs (*e.g. BLAST*, parallelised)
  - Identification of homologous sequences



  - BLAST (heuristic): O(nm) n :length of the query sequence; m: total length of the sequence database (exponential growth)
  - Example (Sun Sparc Ultra-4 300 MHz, 1 Gb RAM) :
    - DNA n=400 m=7 10^9 t= 50'
    - Protein n=172 m=2 10^8 t= 1'20''

- (2) Multiple alignment
- (3) Phylogenetic tree reconstruction

BioInformatic Tools for Comparative Sequence Analysis

- (1) Similarity search programs
- (2) Multiple alignment (*e.g. ClustalW,* heuristic, parallelised)

CLUSTAL W (1.74) multiple sequence alignment

```
Xenopus     ATGCATGGGCCAACATGACCAGGAGTTGGTGTCGGTCCAAACAGCGTT---GGCTCTC
Gallus      ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC-----CAAAATAACACCAACATGCAA
Bos         ATGCATCCGCCAC--ATGACCAGCAGGAGGTAGCACCCAAAACAGCACCAACGTGCA
Homo        ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCA
Mus         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCA--
Rattus      ATGCAT----GCCACCATGACCAGCGGGAGGTAGCTCTCAAAACAGCACCAACGTGCAA
```

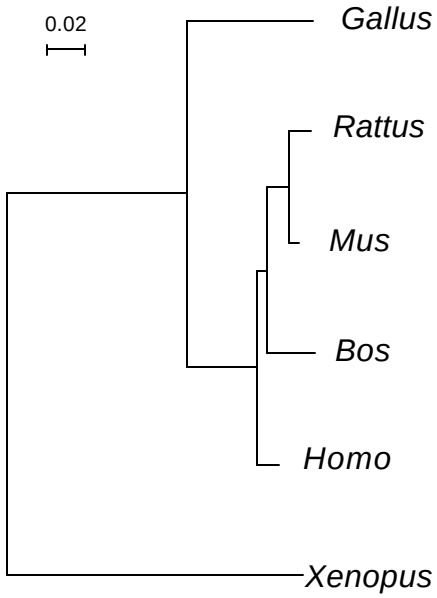
  - Identification of conserved functional elements
  - Phylogenomics (Prediction of gene function)
  - Example: 120 seq. x 300 bases: 18'
- (3) Phylogenetic tree reconstruction

BioInformatic Tools for Comparative Sequence Analysis

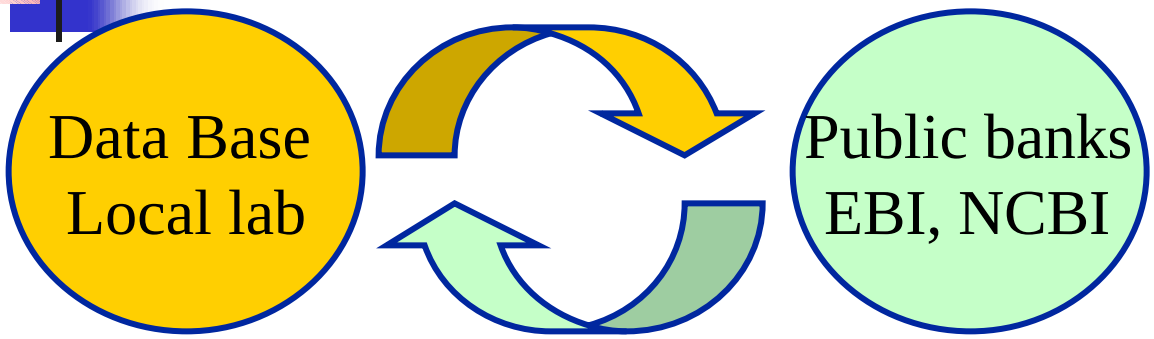
- (1) Similarity search programs
- (2) Multiple alignment
- (3) Phylogenetic tree reconstruction
  - Phylogenomics (Prediction of gene function)

CLUSTAL W (1.74) multiple sequence alignment

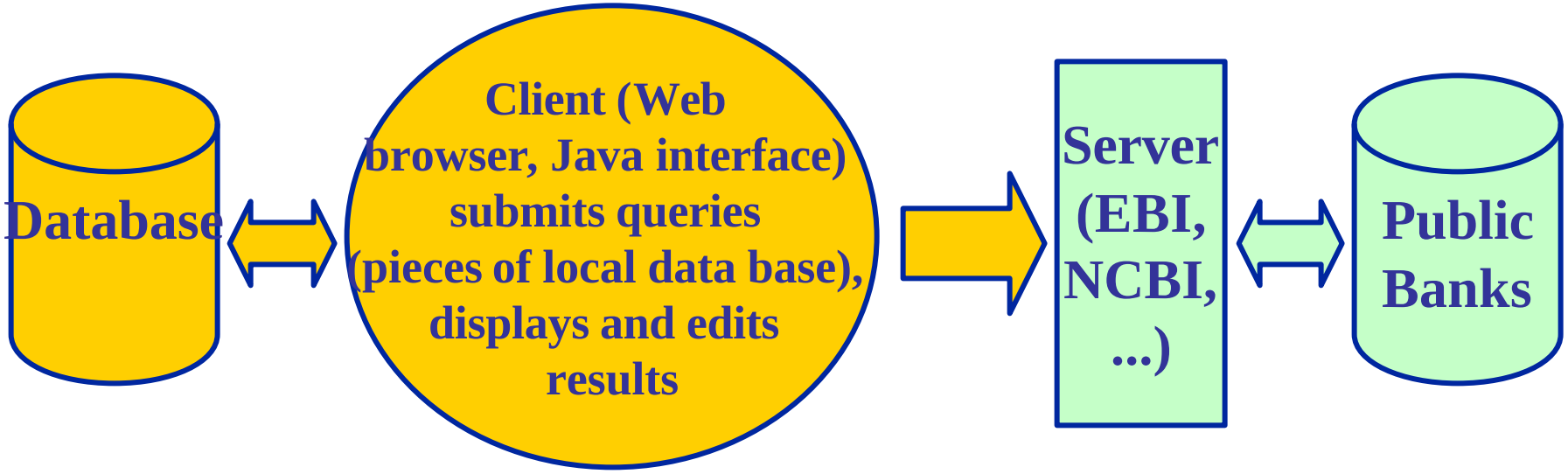
```
Xenopus     ATGCATGGGCCAACATGACCAGGAGTTGGTGTCGG
Gallus      ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC-
Bos         ATGCATCCGCCAC-ATGACCAGCAGGAGGTAGCA
Homo        ATGCATCCGCCACCATGACCAGCAGGAGGTAG
Mus         ATGCATCCGCCACCATGACCAGCAGGAGGTAG
Rattus      ATGCAT--GCCACCATGACCAGCGGGAGGTAGC
```



The typical biologist task : updating a local database



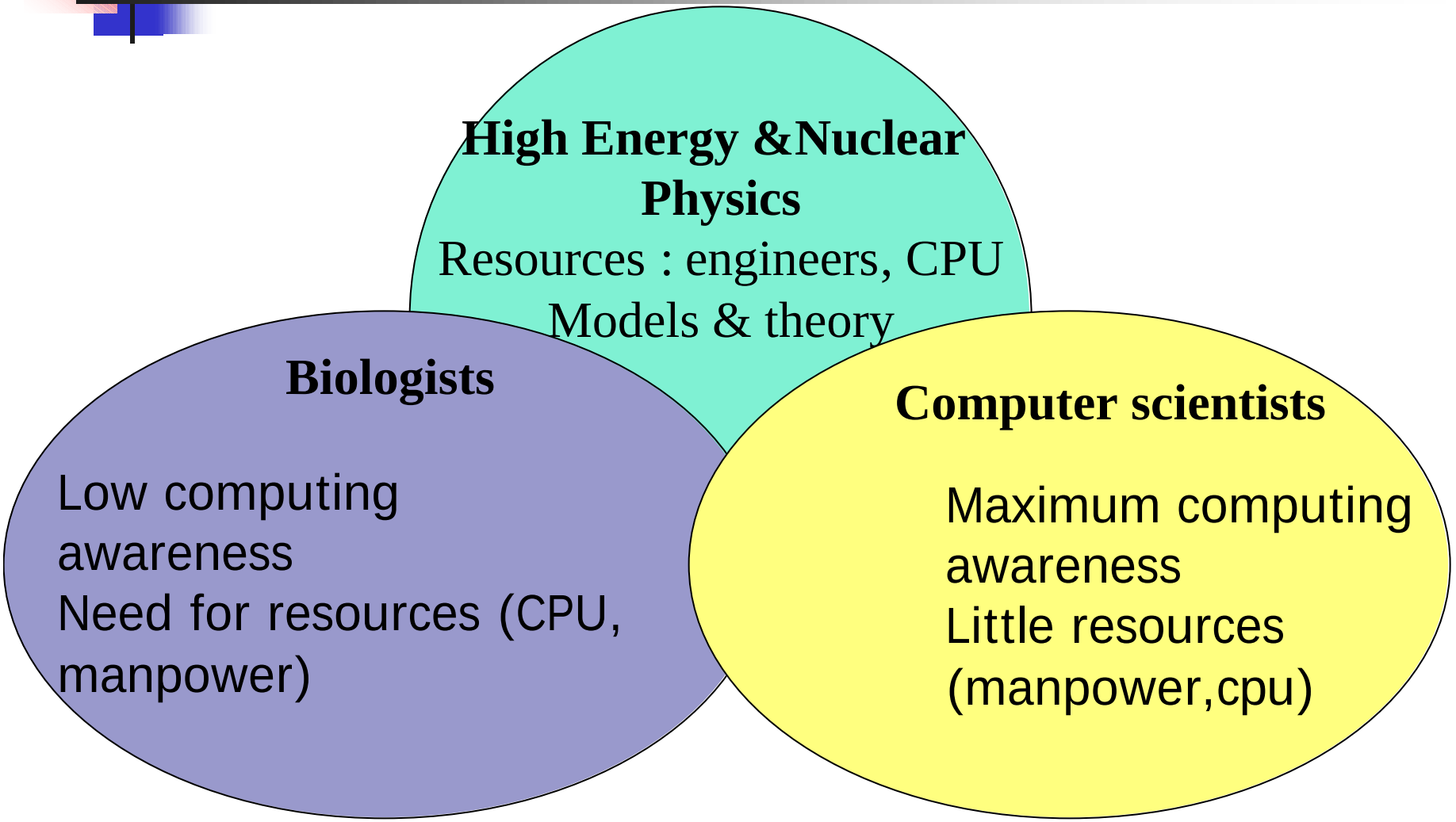
BLAST : search for homologies



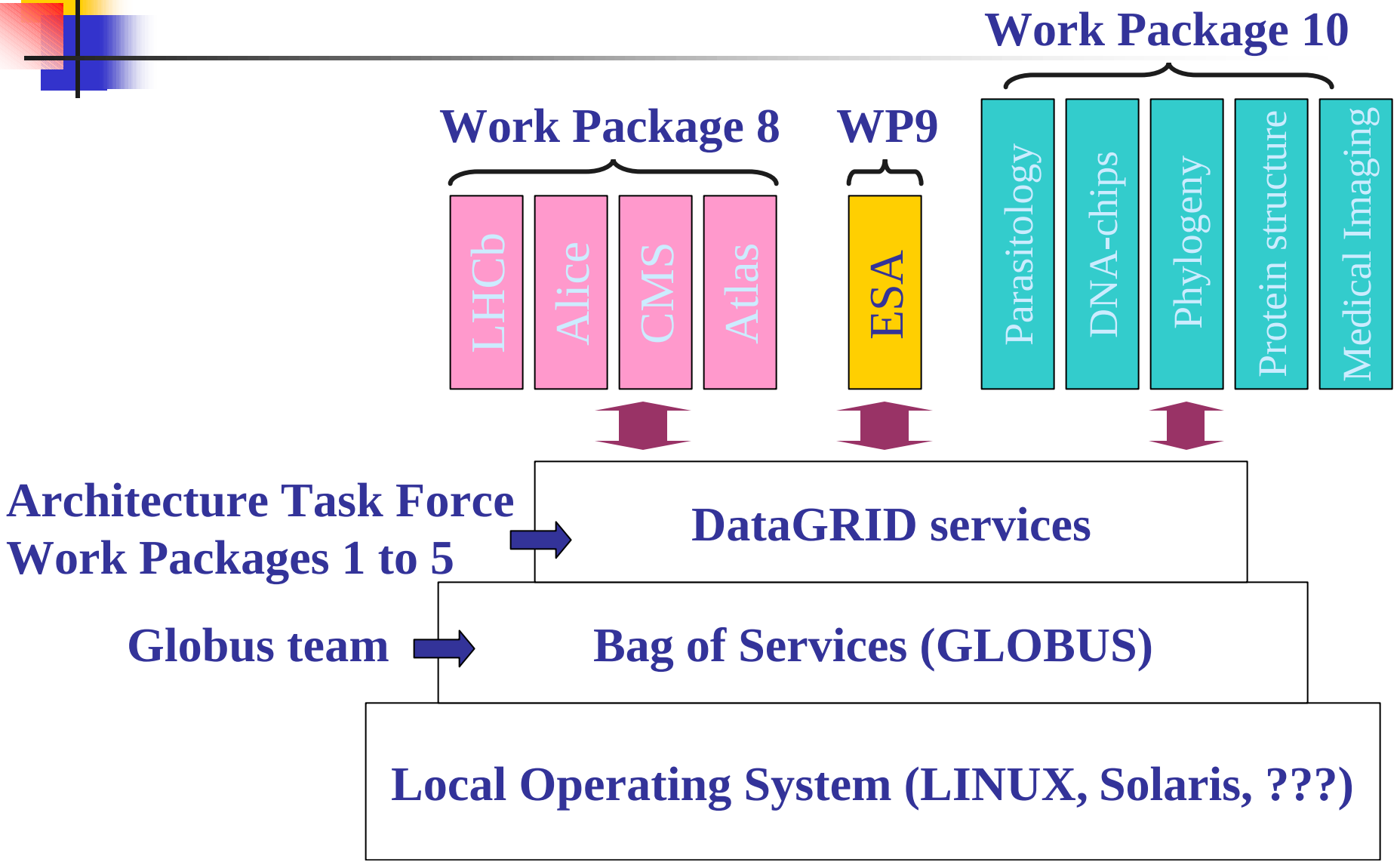
Computer engineering manpower is needed for such a task

- Using remote server CPU
  - Network to access the web server
  - Local database to store the results of the comparison
- To work locally
  - Network to copy the public databases
  - Hard disk to store the data bases
  - CPU to run algorithms

For this task, the biologist needs computing resources



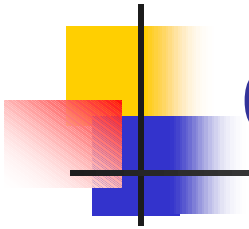
DataGRID, un laboratoire de collaboration pluridisciplinaire



Partners involved in the biology work package

- **CNRS : Clermont-Ferrand, Lyon, Marseille, Montpellier, Orsay, UREC**
- **NFR: Swedish Natural Research Council**
  - **Karolinska Institute, Stockholm**
  - **Uppsala University**
- **University di Padova, Italy**
- **in interaction with ...**
  - **European Bioinformatics Institute**
  - **Institut Pasteur**
  - **INRIA, ...**

DataGRID, un laboratoire de collaboration pluridisiciplinaire



- Opportunité de démontrer les compétences de l'IN2P3 aux autres départements du CNRS
- Mise en commun des compétences et des contacts des laboratoires de l'IN2P3

Conclusion

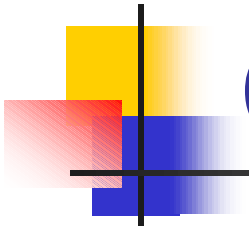
- Le projet de décodage du Génome Humain a été lancé et porté par les physiciens du Department of Energy aux Etats-Unis
- La bioinformatique en France est en retard
- Allons-nous rester hors du coup du post-séquençage ?

Obstacles

- La méfiance des communautés respectives
  - Côté IN2P3 : peur de perdre son âme
  - Côté Sciences du Vivant : crainte vis-à-vis des physiciens
- L'émiettement de la communauté de biologie
  - Chaque laboratoire a sa propre politique informatique
- La difficulté de mettre en place des actions pluridisciplinaires au CNRS

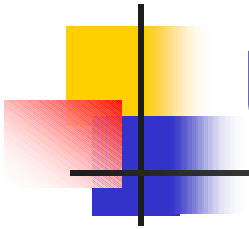
Stratégie appliquée à Clermont

- « Démarchage » des laboratoires de biologie
  - Besoins énormes...
  - Mais très mal quantifiés
  - Et peu d'argent...pour l'instant
- Les collaborations démarrent si une personne est affectée à l'interface entre le LPC et le laboratoire de biologie
  - Stagiaire (bac+2 et au-delà en info)
- Nécessité d'un investissement des personnels permanents pour le suivi



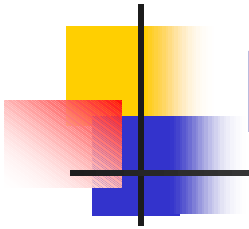
Conclusion (2)

- Les laboratoires de l'IN2P3 peuvent être la colonne vertébrale de la bioinformatique en France
  - 1 labo par région française et par génopôle
  - Des ressources en calcul et en réseau
  - Des compétences en ingénierie informatique
- Prix à payer
  - Aller à la rencontre des biologistes
  - Dédier du personnel temporaire (stagiaires, CDD)
  - Impliquer des personnels permanents



Une proposition

- Mise en place au CCPN Lyon d'un site miroir d'Infobiogen, centre de ressources bio-informatiques français
  - Copie des bases de données
  - Mise à jour hebdomadaire
- Mise en place de sites miroirs du CCPN dans divers labos IN2P3 via DataGRID
- Accès par un portail web à une plate-forme bioinformatique tirant partie de DataGRID



LHC Computing Grid Project

- 250 MSF
- 16 personnes en 2001 à 50 personnes en 2004