

Calcul à hautes performances pour la conception de sondes de puces à ADN et la biologie intégrative

1

□ 2 PROJETS

1. Conception de sondes pour puces à AND
 - Grille EGEE et calcul HPC local
 - Etude du potentiel GPU (Alignement – publi. sept. 2009)
2. Biologie
 - Grille EGEE et calcul hybride – test de grille en local (national Aladdin ?)
(A base de lames proposant 4 GPUs - Couplage machines SMP / GPU)



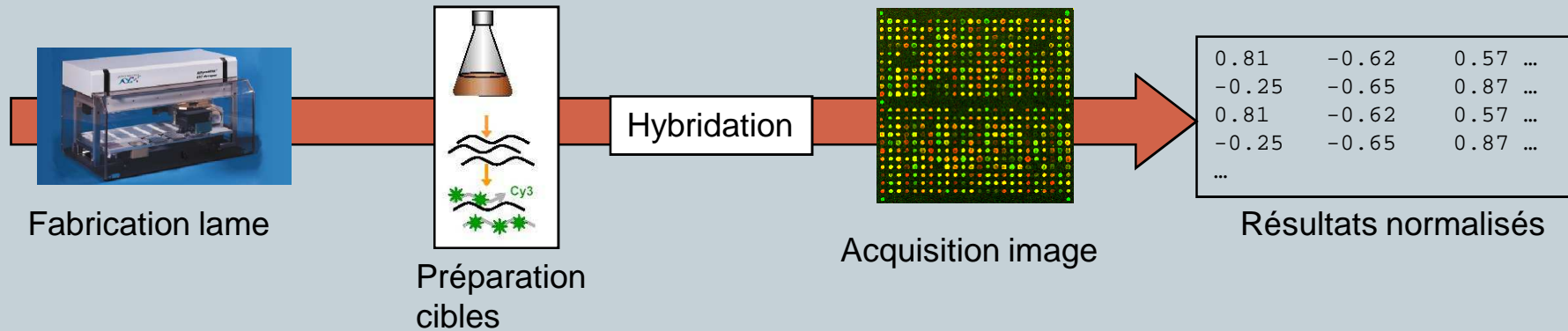
David HILL

David.Hill@univ-bpclermont.fr

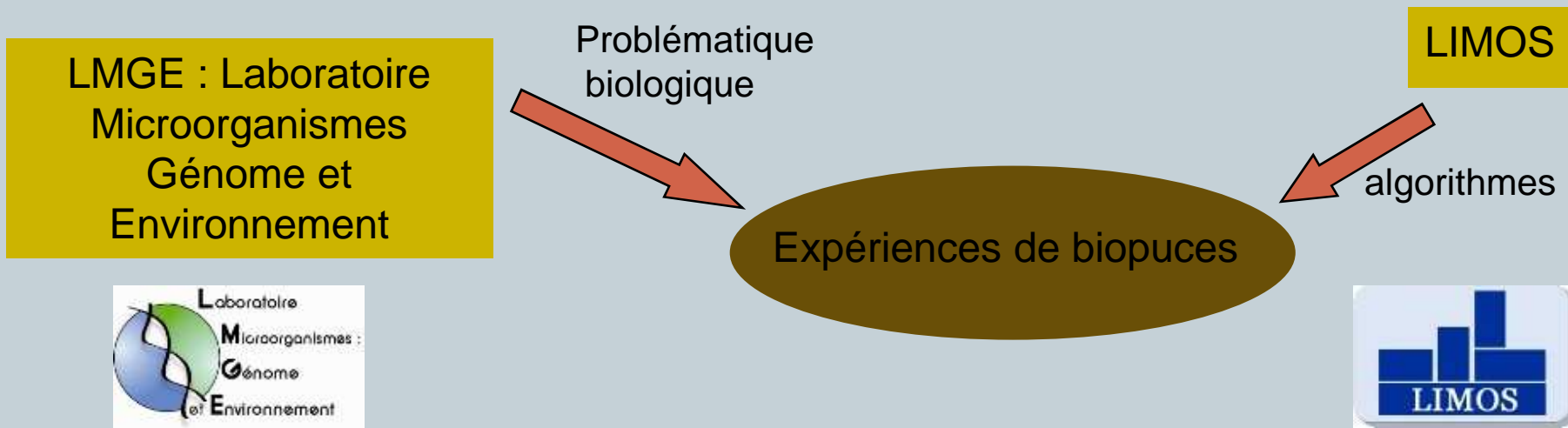


(P1) Recherche sur la conception de sondes (pour puces à ADN)

2



Bioinformatique des puces à ADN

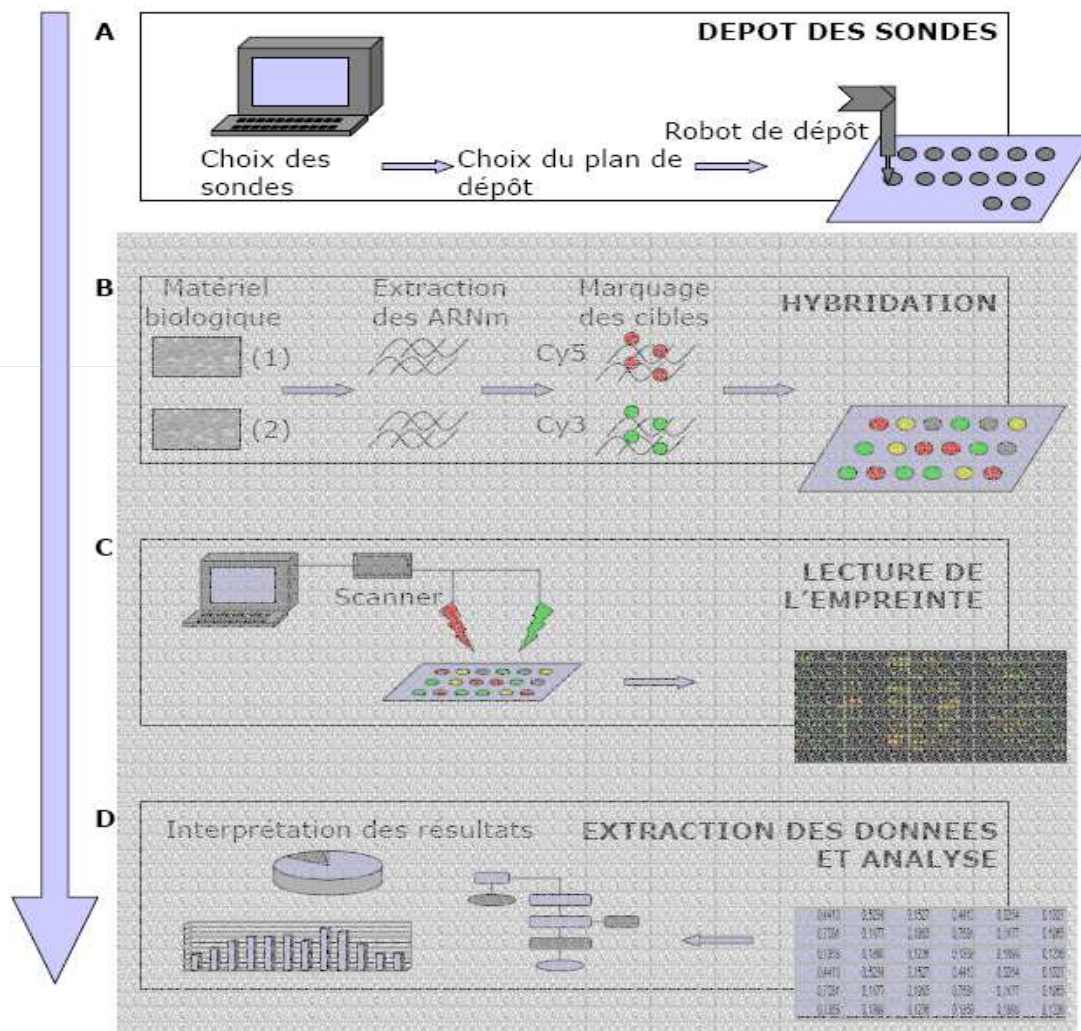


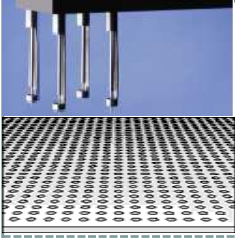
De la conception de sondes à leur validation

3

Conception
« design »
des sondes

Validation
du
« design »





Problématique générale

4

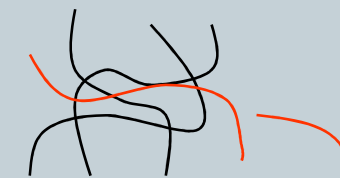
- Conception de méthodes et outils logiciels pour la *conception* de sondes oligonucléotidiques pour puces à ADN
- Application aux biopuces transcriptomiques et aux biopuces de type phylogénétique.
- Développement d'algorithmes pour la recherche de sondes spécifiques
- Traitement de masses de données génomiques

Utilisation de la grille de calcul EGEE

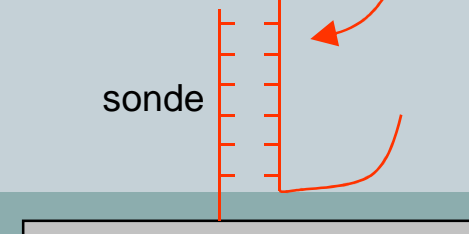
Quelques résultats :

- 2 thèses pluridisciplinaires
- 2 logiciels
- Plusieurs articles dans « Bioinformatics »

mélange cible



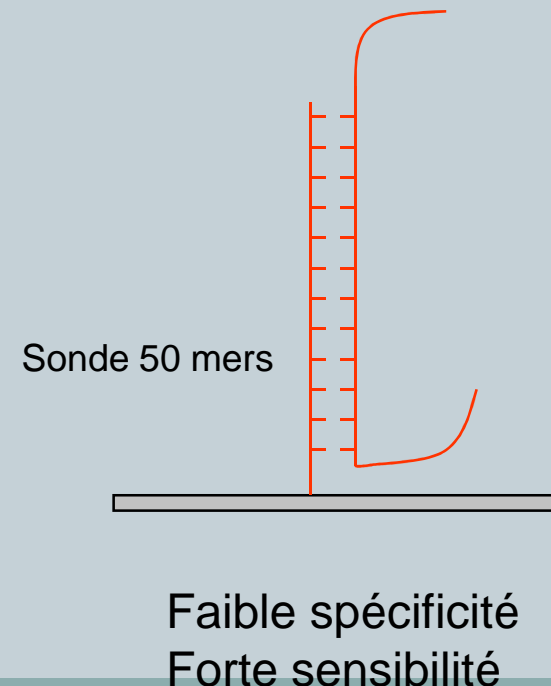
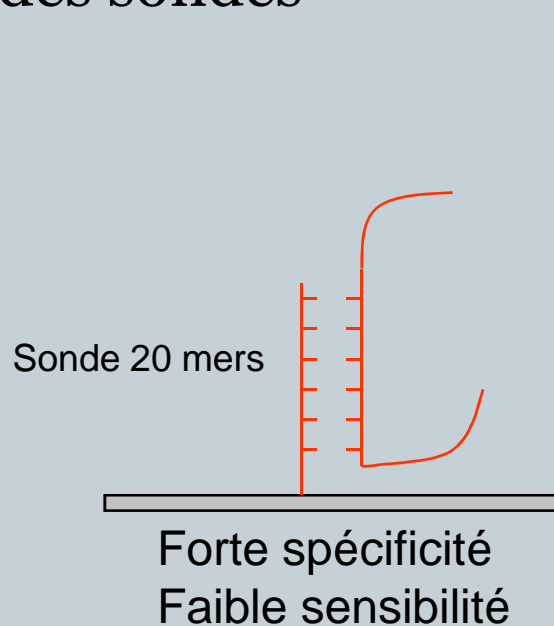
sonde



Problématique « sonde »

5

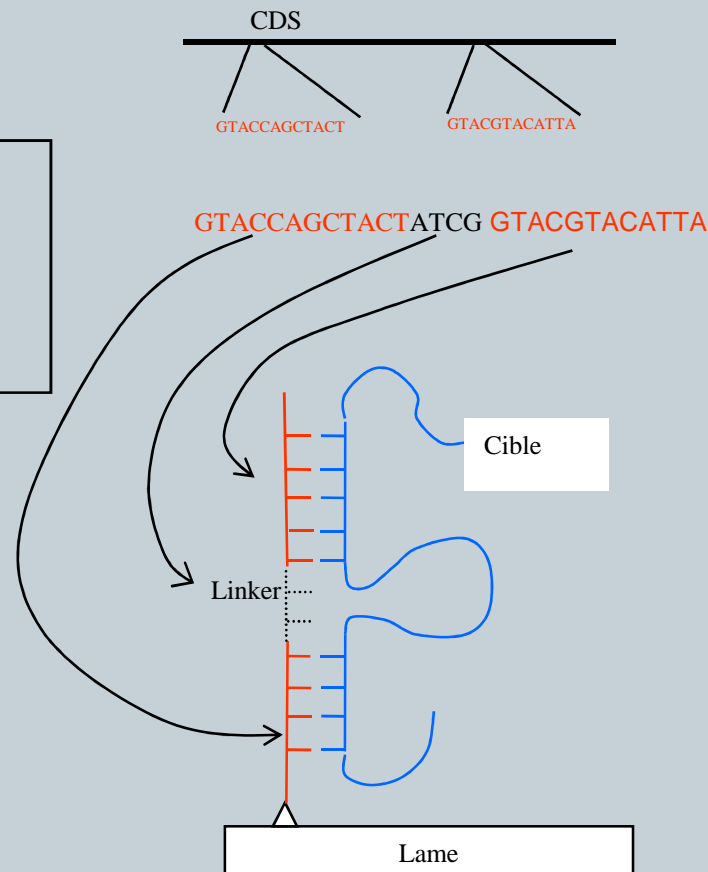
- Problème majeur dans l'approche classique : spécificité et sensibilité des sondes
- Mise au point d'une nouvelle approche pour la conception des sondes



Proposition

6

Approche à :
Forte spécificité
Forte sensibilité



- Applicable dans des milieux biologiques complexes **avec parasites ou virus dans une cellule hôte**
- Validation expérimentale par des biopuces prototypes
- Développement de logiciels



Applications : conception de biopuces phylogénétiques

7

Quelques objectifs applicatifs :

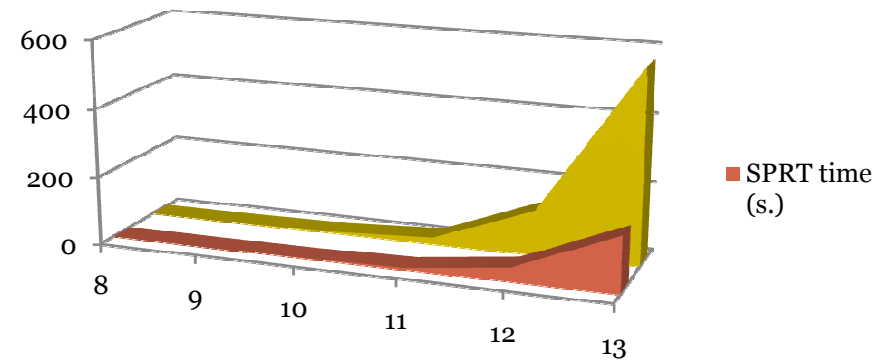
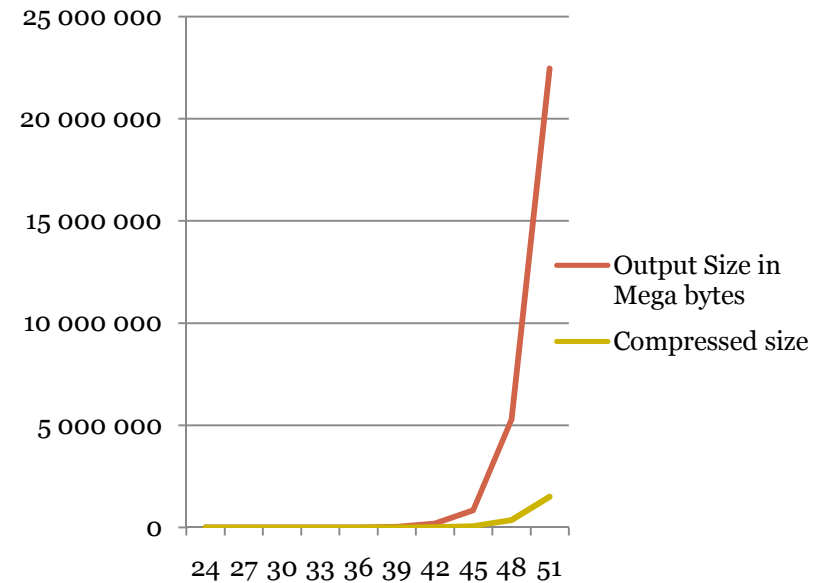
- Etudier la diversité microbienne dans des environnements complexes à l'aide de puces à ADN
- Biorémédiation : approche basée sur l'activité enzymatique des microorganismes
 - Les sondes sont déterminées à partir des protéines intervenant dans des voies métaboliques de **dépollution**
 - **Identification des espèces microbiennes capables de survivre dans les milieux défavorables et participer à la dégradation des polluants**

Découverte de micro-organismes

Masses de données et calculs associés...

8

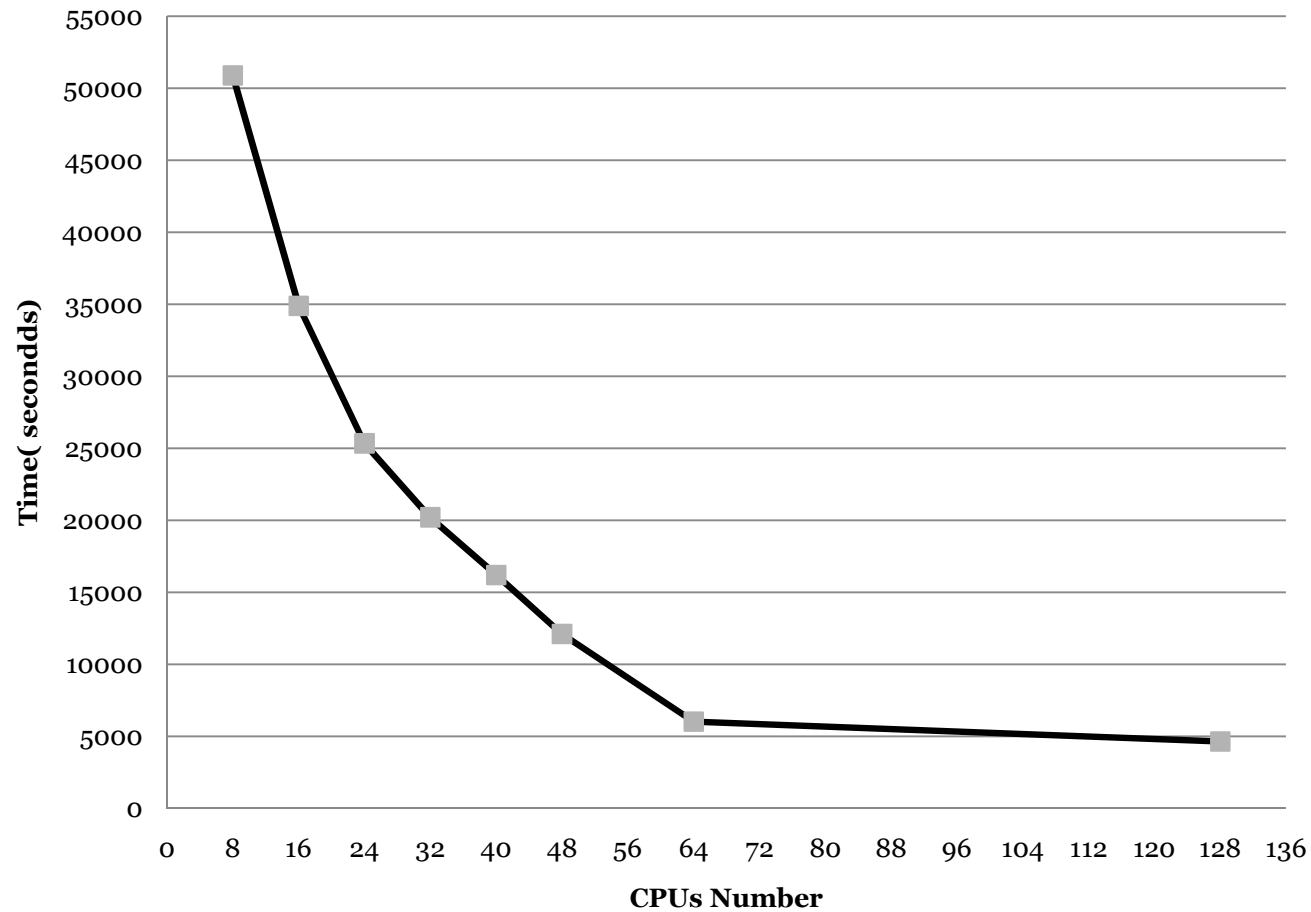
- Traduction inverse complète pour identifier les séquences potentiellement capables de produire les protéines trouvées sur un site
- Proposition d'algorithmes séquentiels puis parallélisation
 - sur fermes de calcul en local
 - puis sur grille
- En cours sur EGEE – calcul des sondes pour tous les procaryotes et champignons connus



Utilisation type MPI-Blast en local (> et test de MPI ...sur “Grille”)

9

MPI-BLAST en local

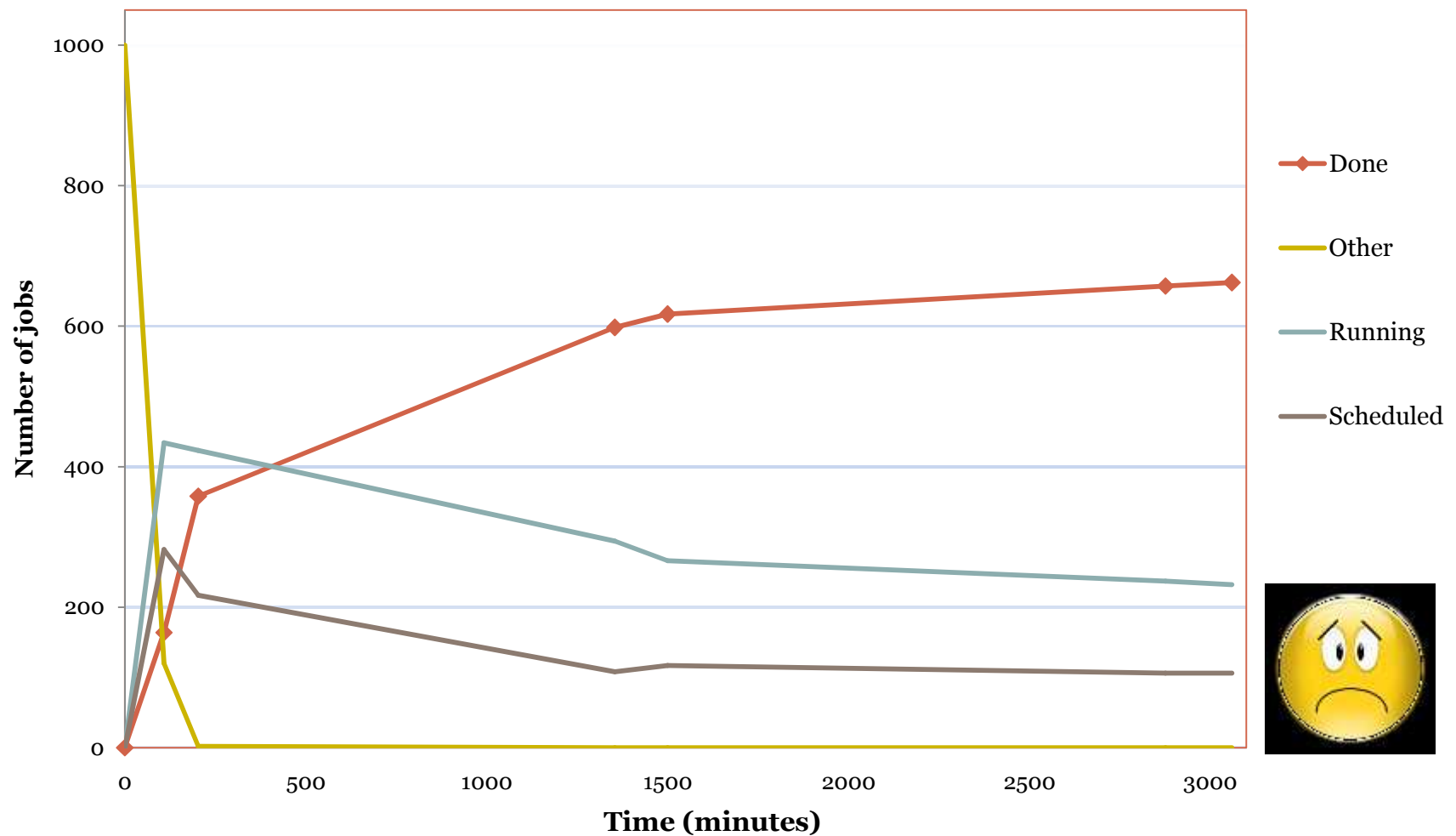


Speedup = ~ 50

Mise au point et performances sur grilles de production

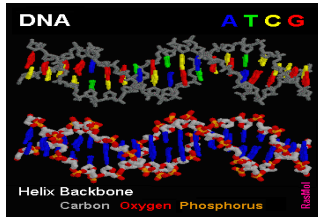
10

« Latency distribution » 1000 jobs avec 20.000 sequences



(P2) Biologie intégrative Multi-Echelle (Sté IBC)

11



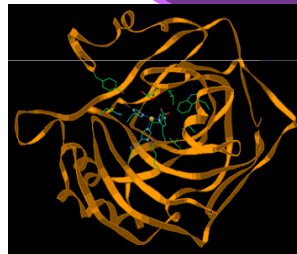
Séquençage du
génome

~30 000 genes in Man

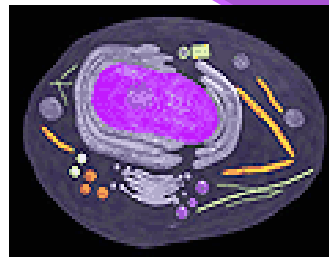
~26 000 genes in Mustard Weed



**Cette complexité ne s'expliquera qu'en
comprenant les réseaux de régulation des
gènes, les voies métaboliques, et
l'intégration des fonctions physiologiques
sur plusieurs échelles**



Modelisation
des
structures
des protéines



Modélisation des voies
métaboliques et de régulation

Modélisation des
organes



(IBC - GMSP) “Generic Modeling and Simulating Platform”

❑ Challenges

La géométrie des tissus vivants est très complexe à toutes les échelles spatiales

Les processus en jeu se déroulent sur des bases de temps très différentes

❑ The GMSP

- **Générateur de modèles multi-échelle**

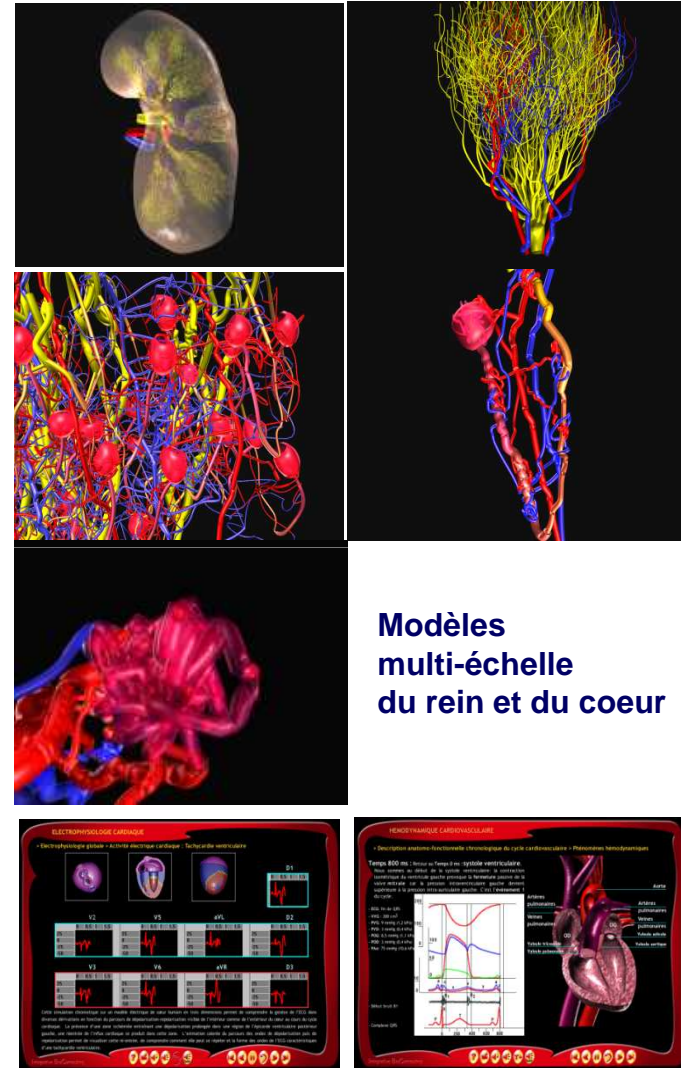
Basé sur les principes d'auto-organisation

- **Connaissances Multi-domaines**

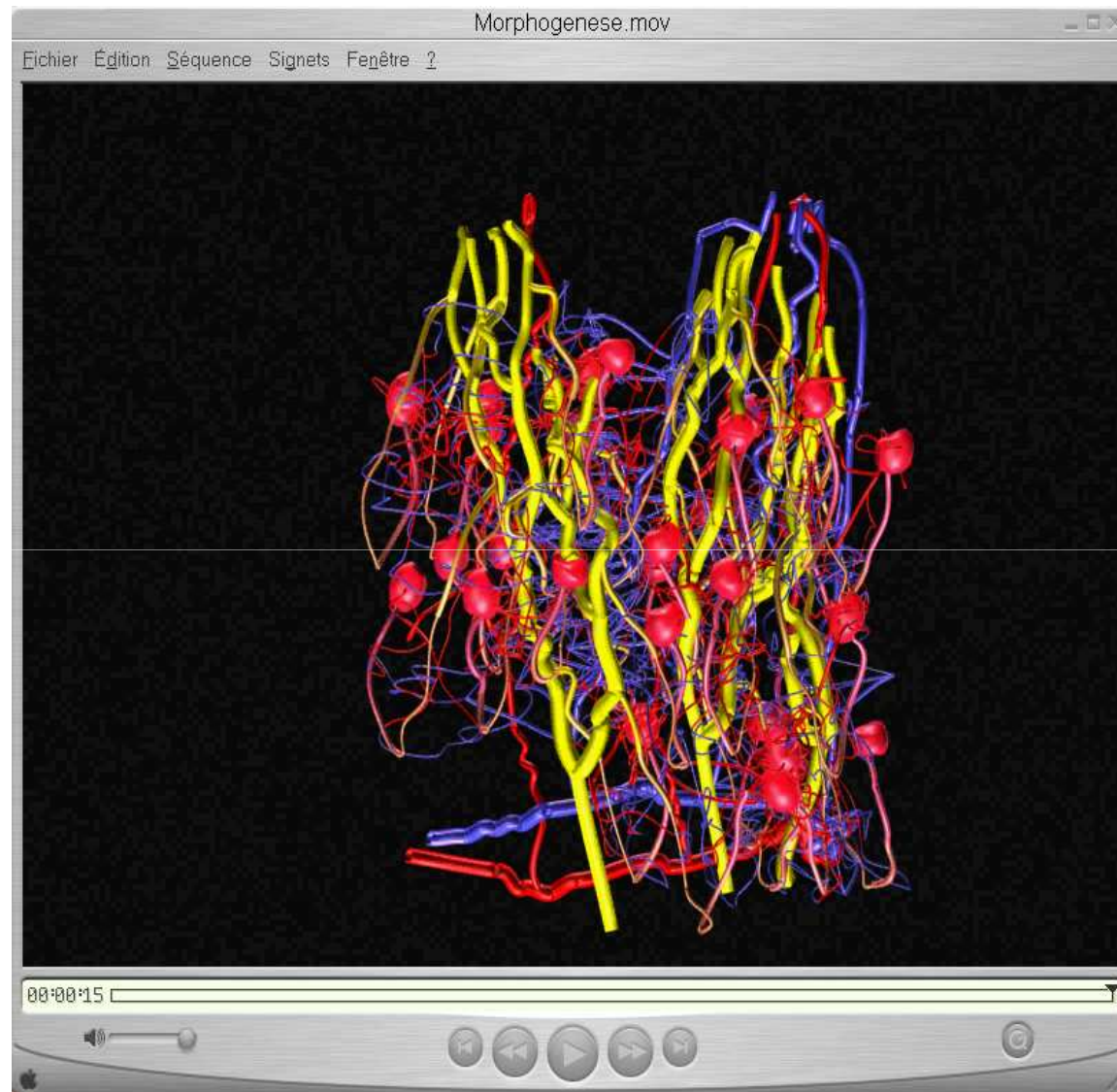
Anatomie, physiologie, ...

- **Modules de raisonnements (IA)**

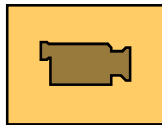
Génération de modèles et de simulation



Génération de modèles par auto-assemblage : morphogenesis

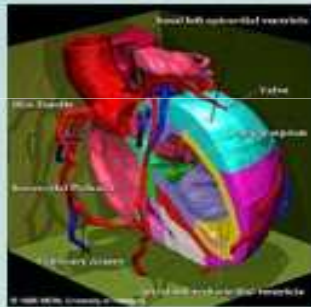


Example : « a trip inside the human body » © IBC

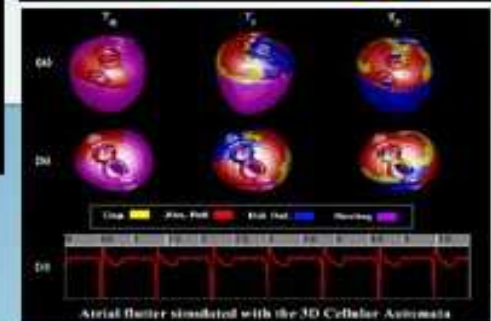
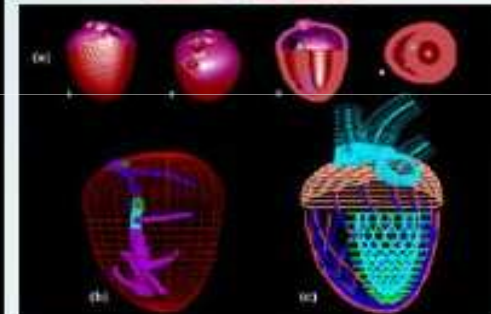


Applications : Education et expérience avec 1 patient virtuel - Aide à la décision

Anatomy

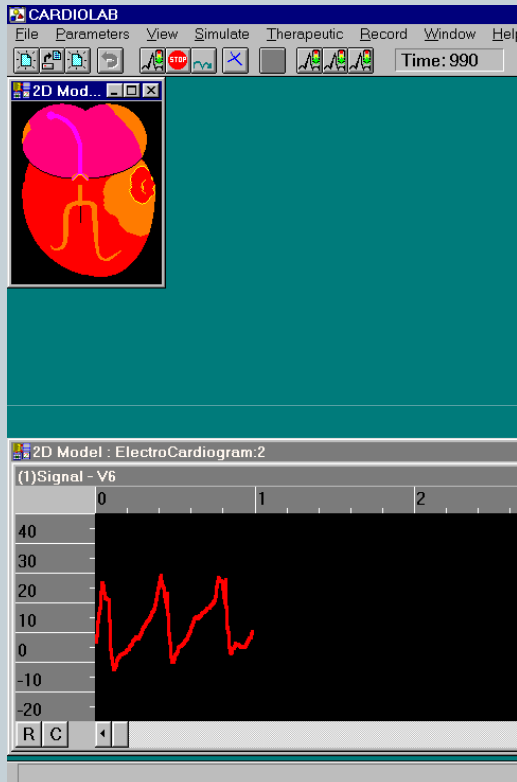


Integrative Physiology

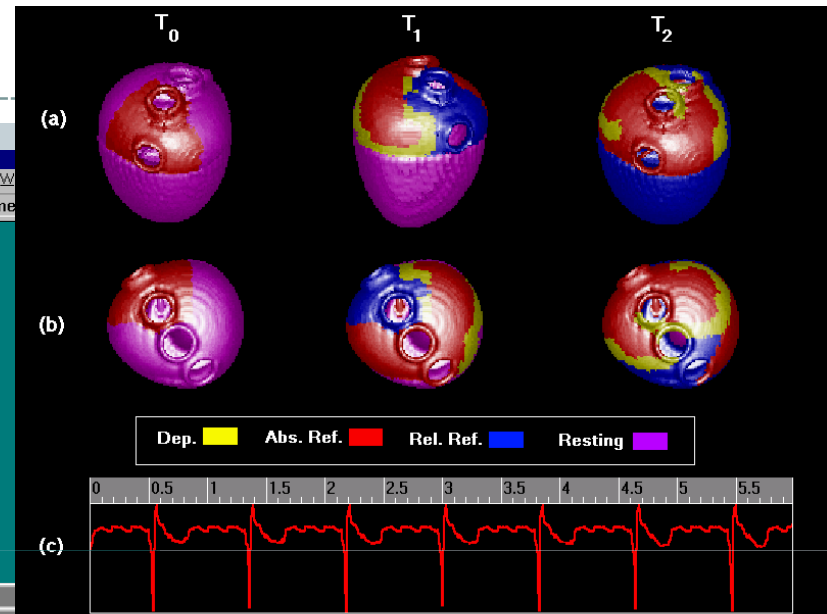
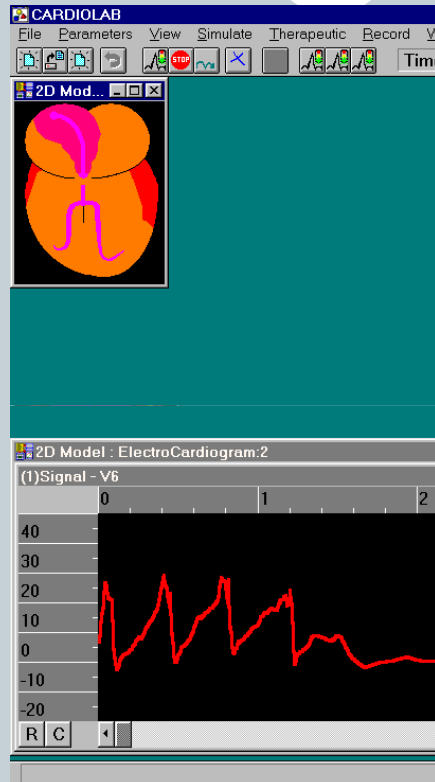


Ex : Cardiolab - test de nouveaux médicaments

16



Problème ventriculaire



Atrial flutter simulated with the 3D Cellular Automata

Après injection de Verapamil

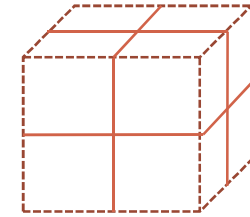
Après injection de Lidocaine

- ❑ Amélioration des méthodes de calcul
- ❑ Manipulation de bases de données importantes
- ❑ Modélisation d’agents “intelligents” pour les organes
 - ❑ e-learning and e-training
 - ❑ e-diagnosis
- ❑ Réalité virtuelle
- ❑ Utilisation de GPU et de grappes de GPU (en grille ?)
- ❑ Utilisation de noeuds pour les besoins conséquents en mémoire RAM
- ❑ Utilisation de la grille de production EGEE
 - ✓ Simulations d’organes
 - ✓ Simulation de soins (Monte Carlo GATE)



BILAN « Application Driven » : besoins en « calcul à hautes performances »

❑ **Division de l'espace 3D (avec des "Octree")**



❑ **Echelle des demandes de ressources IBC**

Scale ratio (X)	N = log ₂ (X) (rounded)	Nb octree cells (rounded)	Nb of PC	
			1 PC	10 ⁶ cells
10 ⁹	30	10 ²⁷	10 ²¹	➔
10 ⁶	20	10 ¹⁸	10 ¹²	
10 ³	10	10 ⁹	10 ³	
10 ²	7	2 · 10 ⁶	2	

❑ **Besoin en mémoire de masse : ~100 Terabytes**

❑ **Quantité de données transmises entre cellules :**

✓ ~ 100 Ko/unité de tps

❑ **Architecture de grilles ?**

✓ **Connexions entre processeurs massivement multicoeurs**

✓ **Au sein d'un même noeud grille**

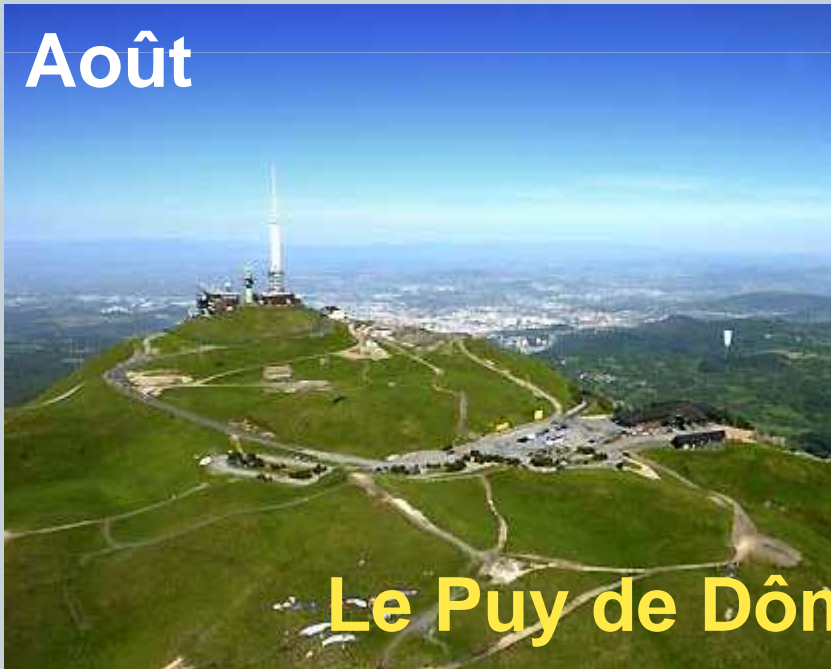
Remerciements :

V. Breton, P. Peyret, M. Missaoui, S. Rimour, R. Reuillon, C. Gouinaud, J. Salzemann, M. Reichstadt, D. Sarramia, J. Caux, P. Siregar (IBC)

19

QUESTIONS ?

Août



Décembre



Le Puy de Dôme - Clermont-Ferrand