LIMOS
UMR 6158 CNRS

Rencontres scientifiques France Grilles
Lyon, le 19 septembre 2011

Laboratoire Microorganismes : Génome et Environnement

Détermination de sondes oligonucléotidiques pour biopuces phylogénétiques en environnement grille de calcul

F. Jaziri*, S. Cipière*, M. Missaoui*, J. Denonfoux, E. Dugat-Bony, N. Parisot, A. Mahul, S. Rimour, E. Peyretaillade, P. Peyret*, D. Hill*

* Co premier auteur / Co directeur

Présenté par : Faouzi Jaziri

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP
FranceGrilles
AuverGrid
CONSEIL RÉGIONAL AUVERGNE
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

✓ L'étude des **micro-organismes** dans des **milieux complexes**.

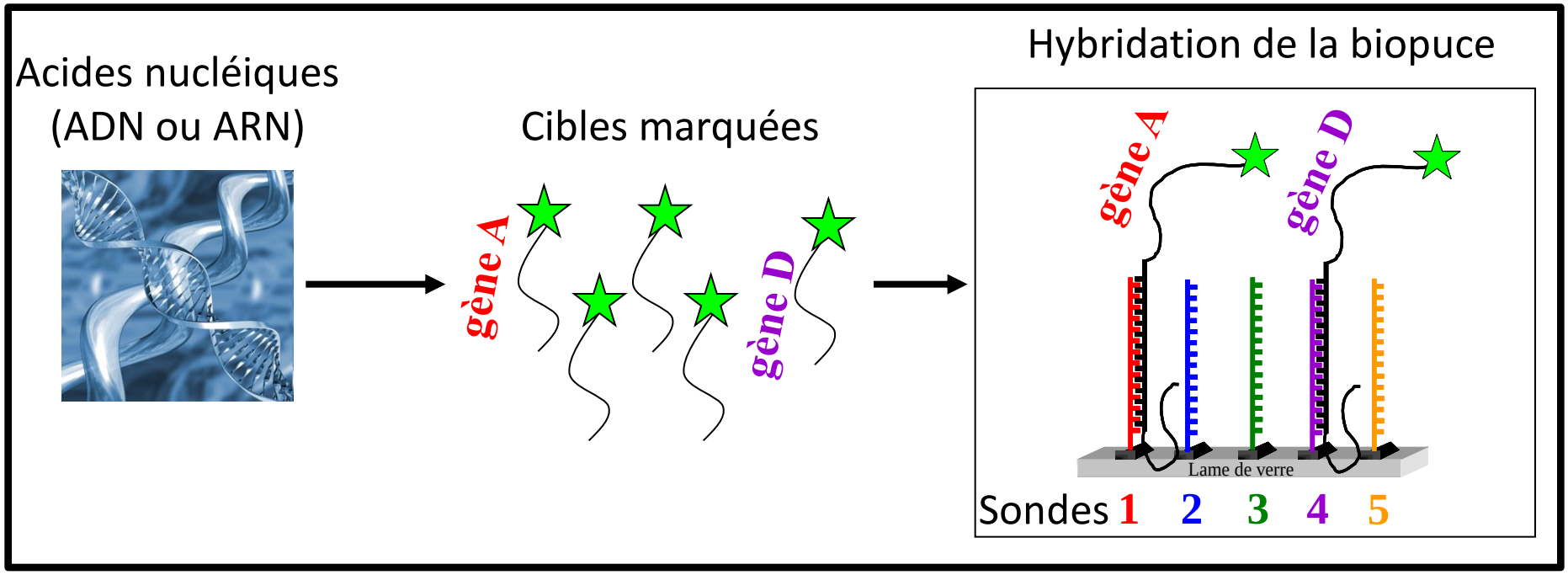
✓ 14 millions d'espèces vivantes sur notre planète, moins de 2 millions ont été décrites scientifiquement (Dunbar et al., 1999) : la plupart des espèces vivantes sont des micro-organismes.

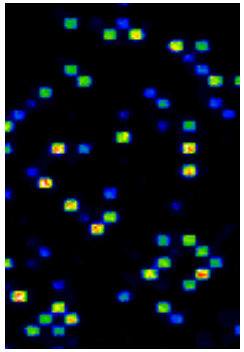
✓ Les **micro-organismes** constituent la plus grande diversité du monde vivant : acteurs essentiels au bon fonctionnement des **écosystèmes**.

✓ La compréhension du fonctionnement des écosystèmes et les rôles des micro-organismes : un enjeu majeur de l'**écologie microbienne**.

➡ Le besoin de développer des nouvelles **approches haut débit** de génomique permettant d'appréhender la complexité des écosystèmes.

Biopuces ADN : outils à haut débit de choix permettant l'étude du niveau d'expression des gènes ou la présence de plusieurs milliers d'espèces en une seule expérience (Schena et al., 1995).

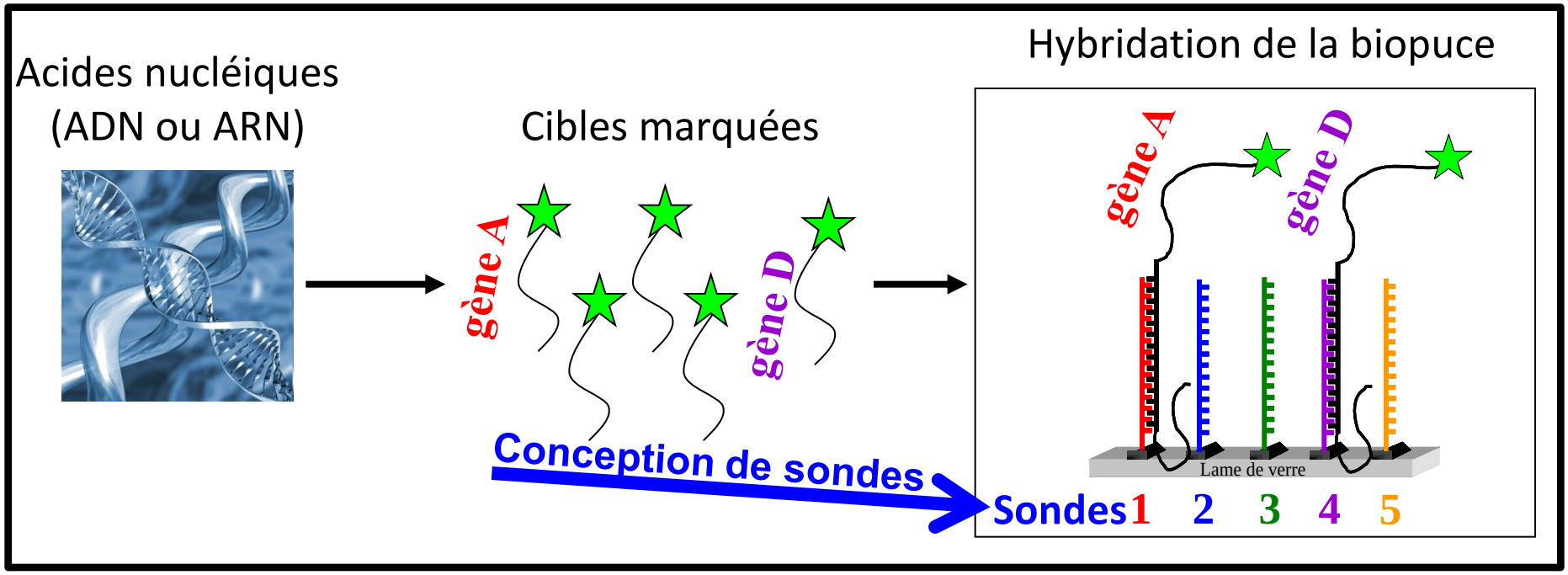


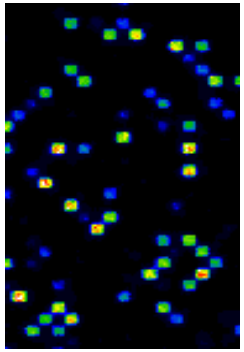


Formats actuels : jusqu'à 4,2 millions de sondes

Suivi de l'expression de centaines de milliers de gènes en une seule expérience

Biopuces ADN : outils à haut débit de choix permettant l'étude du niveau d'expression des gènes ou la présence de plusieurs milliers d'espèces en une seule expérience (Schena et al., 1995).





Formats actuels : jusqu'à 4,2 millions de sondes

Suivi de l'expression de centaines de milliers de gènes en une seule expérience

✓ **Conception de sondes :** étape complexe mais cruciale dans la conception de biopuces à ADN (Kreil et al, 2006).

✓ Sélection de sondes **sensibles** et **spécifiques**.

✓ Recherche de sondes **spécifiques** : comparaisons de séquences pour la **recherche de similarité :**

↑ nombre de séquences déposées dans les bases de données internationales

↓

↑ nombre de comparaisons pour la recherche de similarité

→ **La recherche de similarité pour la conception de sondes :** tâche très lourde en temps de calcul (Wu and Tseng, 2005) .

→ Accroissement continue des besoins de la conception de sondes en termes de calcul.

Exemple : CommOligo (Li et al., 2005), ROSO (Reymond et al., 2004) ou PhylArray (Militon et al., 2007) peuvent nécessiter plusieurs heures pour la sélection de sondes pour un seul gène.

➡ Développement d'un algorithme de conception de sondes pour biopuces phylogénétiques, entièrement déployé sur la grille de calcul européenne EGI (European Grid Infrastructure).

✓ L'algorithme est basé sur nos développements antérieurs (Militon et al., 2007).

➡ Objectif :

✓ Une meilleure sélection de sondes en terme de sensibilité et de spécificité.

✓ Une meilleure performance avec l'utilisation de la grille de calcul pour la distribution du calcul à grande échelle.

✓ Implémentation : C++ et Perl.

✓ Algorithme de sélection de sondes déployé sur la grille de calcul européenne.

✓ Critères de sélection des sondes oligonucléotidiques : spécificité, taille, taux de dégénérescence, et nombre maximum d'hybridations croisées autorisées.

✓ fonctionnalité : conception simultanée de milliers de sondes spécifiques et exploratoires pour l'étude de la présence et/ou l'évolution des procaryotes et champignons dans différents milieux complexes.

✓ Les principales étapes de l'approche proposée:

1. **Construction d'une base de données de séquences procaryotes et fongiques.**
2. **Sélection de sondes.**

1. Construction de la bases de données :

- ✓ Une base de données de séquences procaryotes et fongiques de petites sous unités de l'ARN Ribosomique (rRNA SSU) : 16S pour les procaryotes et 18S pour les champignons.

- ✓ Les différentes étapes de l'algorithme de construction de la base de séquences :

  a. Extraction des séquences de la base EMBL.

  b. Suppression des séquences de mauvaise qualité.

  c. Elimination des séquences redondantes à l'aide de BlastClust.

  d. Vérification et correction de l'orientation des séquences.

  e. Construction de sous groupes homogènes de séquences et suppression des séquences mal annotées en utilisant la technique de classification « k-means » et des versions modifiées de l'algorithme Clustalw.

2. Conception de sondes oligonucléotidiques :

- ✓ L'algorithme de sélection de sondes consiste en plusieurs étapes (Figure2).

- ✓ Première étape : construction des séquences consensus à partir des alignements (Clustalw_MPI) des groupes de séquences obtenus après construction de la base SSU rRNA : **traitement lourd pour les gros groupes de séquences**.

2. Conception de sondes oligonucléotidiques :

✓ L'algorithme de sélection de sondes consiste en plusieurs étapes (Figure2).

✓ Première étape : construction des séquences consensus à partir des alignements (Clustalw_MPI) des groupes de séquences obtenus après construction de la base SSU rRNA : **traitement lourd pour les gros groupes de séquences**.

➡ **Alignement d'alignements** pour les gros groupes de séquences : un groupe est partagé en sous groupes de séquences à l'aide de Blastclust. Les séquences des sous groupes sont alignées à l'aide de Clustalw-MPI. Ces alignements sont ensuite alignés à l'aide du logiciel Opal ➡ **meilleure performance en temps de calcul (tableau1) :**

Groupes alignés	Nombre de séquences	Nombre de sous groupes	Temps d'alignement (secondes)	
			Clustalw-MPI	Alignement d'alignements
Pichia	128	5	320	225.6
Vibrio	1174	37	2542	1247
Bacillus	3947	58	12586	3130

Tableau 1. Performance de l'alignement parallèle et d'alignement d'alignements en utilisant en local 100 processeurs sur un cluster de calcul.

- ✓ Deuxième étape : la plus coûteuse en termes de temps de calcul : le test de spécificité de toutes les sondes possibles pour déterminer leurs hybridations croisées potentielles. **Cette partie est entièrement déployée sur la grille de calcul européenne.**

- ✓ **Déploiement sur grille** : découpage de la séquence consensus en plusieurs fragments (Figure1). Le nombre de fragments est variable d'un design à un autre. Il est déterminé en fonction d'une estimation préalable du nombre de sondes générées.

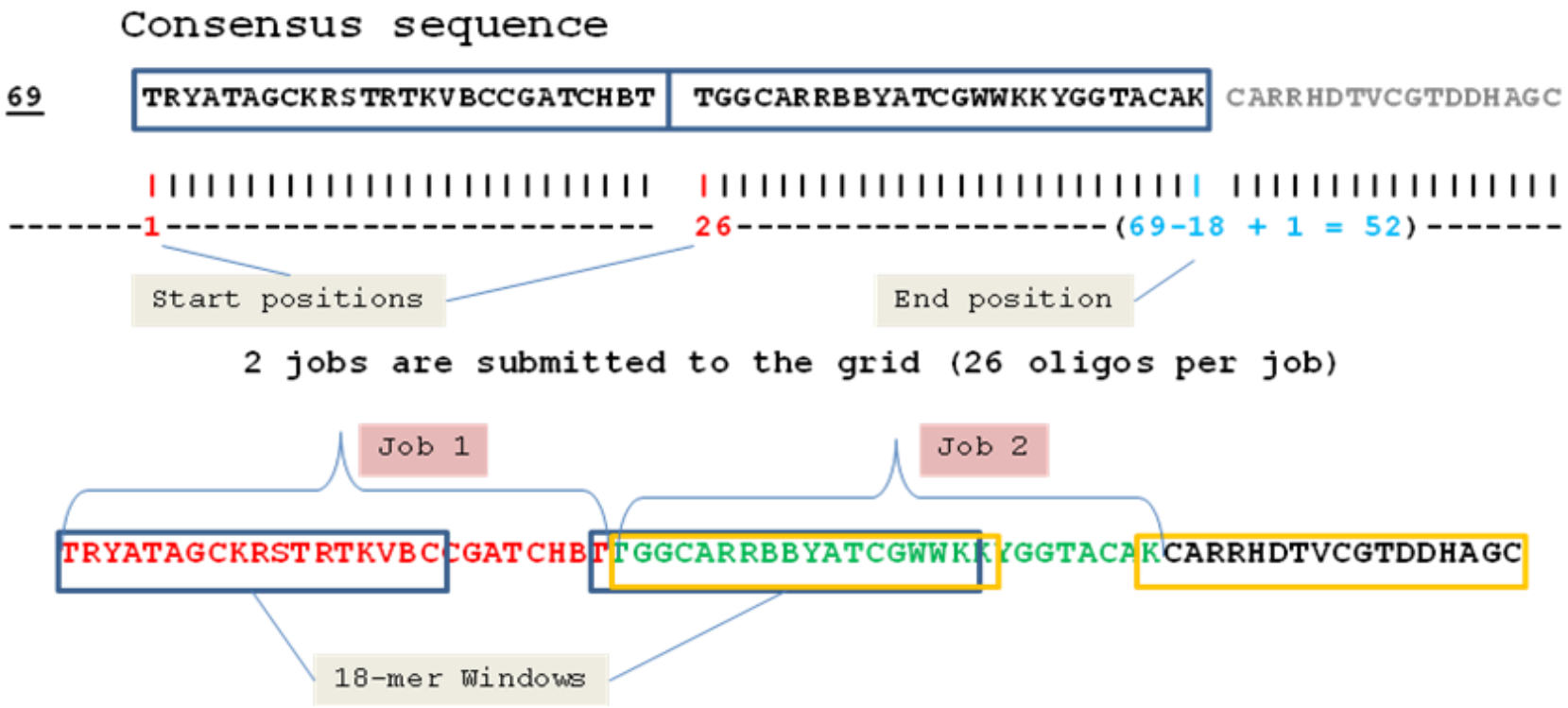


Figure 1. Fragmentation de la séquence consensus pour soumission à la grille

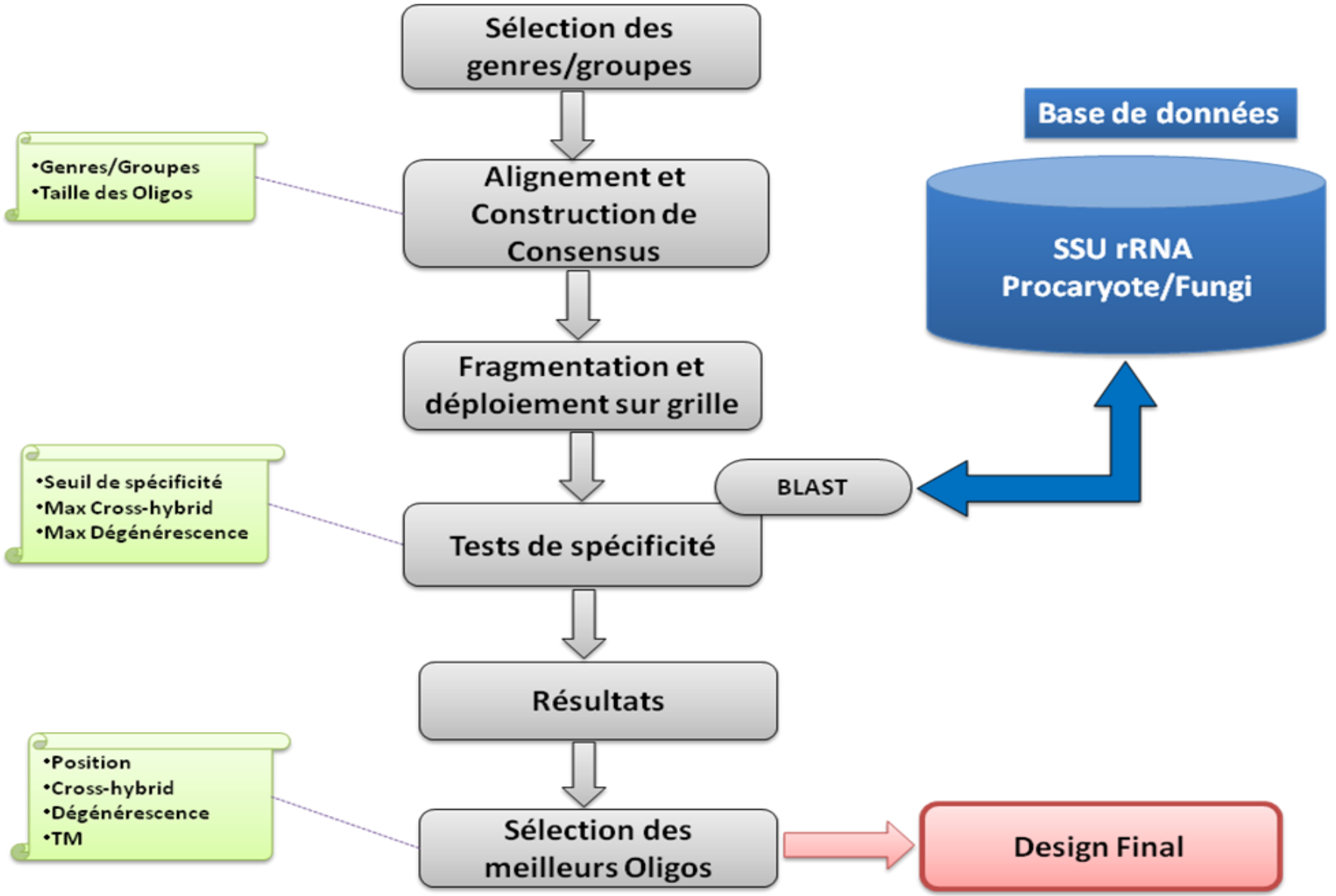


Figure 2. Les différentes étapes de l'algorithme de conception de sondes proposé

- ✓ Conception de sondes oligonucléotidiques de haute qualité, spécifiques et exploratoires pour un total de **2072 genres procaryotes et 1441 genres fongiques**.

- ✓ Test de performance en termes de temps de calcul : conception de sondes pour tous les genres fongiques obtenus après la création de la base ARN ribosomique.

- ✓ Durée moyenne par conception sur une seule CPU : 6 heures et demi.

➡ Temps de calcul total estimé pour les champignons sur une seule CPU : un peu **plus d'un an-CPU**.

- ✓ Seuls les jobs terminés avec succès dont les résultats finaux ont été obtenus sont comptabilisés.

- ✓ Après **7 semaines** **de calcul sur la grille européenne**, nous avons obtenu les résultats pour 1356 genres des champignons : environ 95% des données (Figure3). Les 5% restants ont été resoumis dans un deuxième temps.

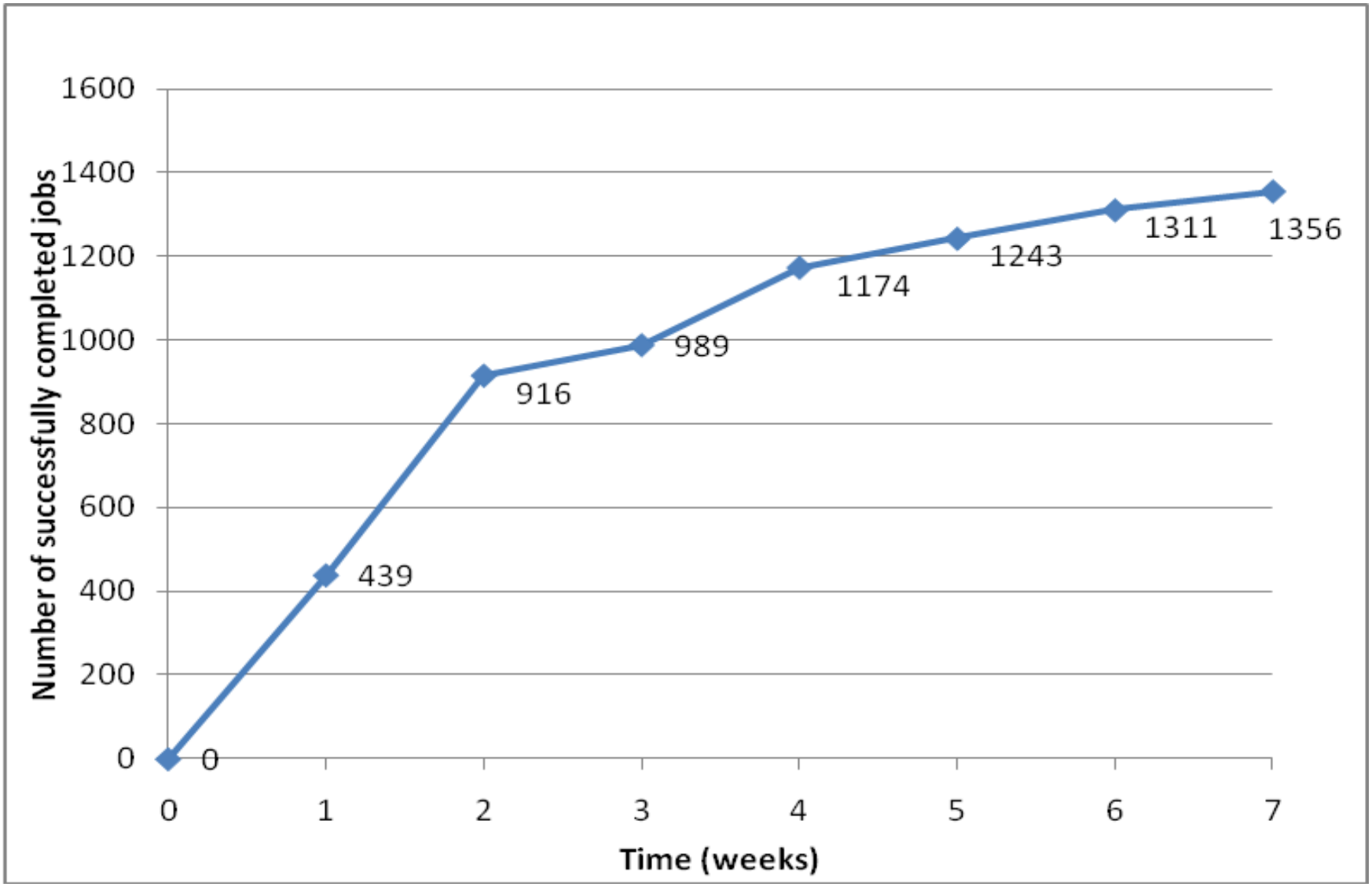


Figure 3. Résultats de calcul pour 1441 genres fongiques sur la grille EGI

Conclusion :

- ✓ Approche proposée : **détermination de sondes** oligonucléotidiques pour biopuces phylogénétiques procaryotes et fongiques **en environnement grille de calcul**.

➡ L'algorithme développé permet de **concevoir simultanément des milliers de sondes spécifiques et exploratoires** pour l'étude de la présence et/ou l'évolution des procaryotes et champignons dans différents milieux complexes.

- ✓ Les performances sont dépendantes de l'état de la grille de calcul.

- ✓ Notre logiciel reste très fiable lorsque la grille est disponible. Il faut faire face à des problèmes lié à la grilles : indisponibilité de certaines ressources de la grille, blocage d'un job au niveau d'un CE, incidents au niveau du réseau.

Perspectives :

- ✓ **Performance** : technique des **« Pilot Jobs »**, **DIRAC**

- ✓ **Accessibilité** : Création d'une plateforme web qui permet à toute la communauté d'utiliser notre logiciel et l'ensemble des logiciels de bioinformatique développés ces dernières années par notre équipe.

LIMOS
UMR 6158 CNRS

Laboratoire Microorganismes : Génome et Environnement

Journée de sensibilisation au calcul intensif en biologie intégrative et écologie

Bordeaux, le 14 septembre 2011

Merci de votre attention

Questions?

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

FranceGrilles

AuverGrid

CONSEIL RÉGIONAL AUVERGNE

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

References

✓ O.Wilson E. et al. (1988), *Biodiversity*, National Academy Press, march.

✓ Dunbar J. et al. (1999), Levels of of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. Appl. Environ. Microbial., 33. D294-296.

✓ Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, 270, 467-470.

✓ Kreil, D.P., Russell, R.R. and Russell, S. (2006) Microarray oligonucleotide probes, Methods in enzymology, 410, 73-98.

✓ Wu, X. and Tseng, C.-W. (2005) Searching Sequence Databases Using High Performance BLASTs In Sons, J.W. (ed), Parallel Computing for Bioinformatics and Computational Biology.

✓ Militon C. et al., (2007), PhylArray: phylogenetic probe design algorithm for microarray, Bioinformatics (2007) 23 (19): 2550-2557.

✓ Li, X., He, Z. and Zhou, J. (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation, *Nucleic Acids Research,* ***33, 6114-6123.***

✓ Reymond, N., Duret, H.C.L., Calevro, F., Beslon, G. and Fayard, J.-M. (2004) ROSO: optimizing oligonucleotide probes for microarrays, Bioinformatics, 20, 271-273.