SeBIG

(Projet de Service 2004)

Service de Bio-informatique de l'IGBMC

http://bips.u-strasbg.fr/

CONTEXTE

En réponse à la production massive de données issues des grands programmes de séquençage, de transcriptomique, de protéomique, etc. la bio-informatique est entrée de plain-pied dans l'ère du haut débit. Le traitement automatisé de l'information biologique nécessite de grandes capacités de stockage et de calcul que la mutualisation des ressources informatiques peut rendre accessibles au plus grand nombre.

De plus, face à des problématiques biologiques de plus en plus complexes, la mise au point de procédures d'analyse automatisées se heurte à deux écueils principaux: (i) le bruit engendré par la grande quantité de données hétérogènes disponible et qu'il faut savoir filtrer pour en extraire le "signal" adapté à la question posée; (ii) la qualité de l'information : les informations produites de manière automatique et intégrées dans les banques contiennent parfois des erreurs et il faut être capable d'estimer la qualité des données et des résultats.

OBJECTIFS

Développement du service de bio-informatique (site d'Illkirch, IGBMC) qui rassemble sur un même site les ressources bio-informatiques (banques, logiciels et accès) de domaines scientifiques permettant l'étude du gène à sa fonction.

Ce service comprend la maintenance et la mise en œuvre de ressources généralistes et thématiques (banques, logiciels) à l'usage des biologistes locaux, nationaux et internationaux; l'adaptation au haut-débit (technologie GRID, procédés automatisés de filtrage et de validation des données,…); la formation, l'assistance et le soutien scientifique pour l'utilisation des outils, la mise au point de protocoles et méthodologies en bio-analyse et bio-informatique, l'interprétation des résultats et l'accès structuré aux données.

PROJET

Le service bio-informatique de l'IGBMC ne vise pas seulement la mise à disposition de moyens et d'outils (programmes, moyens de calculs, banques de données,..) mais également à réunir des compétences et des outils pluridisciplinaires – bio-informatiques, algorithmiques et biologiques – au service du développement de programmes scientifiques nouveaux.

Dans le cadre du présent projet, la plate-forme procédera à la mise en place de protocoles automatisés d'extraction d'information (PipeAlign, DbW,…) pour la bio-informatique à haut débit qu'elle déploiera dans le cadre de ressources mutualisées (technologie GRID,…). Ces outils seront en outre associés à des méthodes de validation automatiques ou semi-automatiques (LEON, NorMD, vALId) permettant d'évaluer la qualité des résultats obtenus.

RESULTATS

- Mise à jour semi-automatisée des banques avec validation hors ligne
- **DbW**, logiciel de veille automatique pour la mise à jour d'alignements multiples de protéines spécifiques de familles fonctionnelles. [2]
- **vALId**, logiciel Web permettant de valider et d'évaluer la qualité de prédiction de séquences de protéines à partir de l'examen automatique d'alignments multiples de séquences complètes [1]
- **LEON**, logiciel de prédiction de l'homologie entre protéines à partir d'alignements multiples [3]
- **Services Web** (SRS, EMBOSS, PipeAlign, GoANNO,…) totalisant plus de 6000 visiteurs par mois venant de 2000 sites différents:
  PipeAlign a été en avril 2005 le point d'entrée sur le site pour plus de 1000 visiteurs.
  Après trois mois de disponibilité sur le site, le service DbW assure automatiquement la mise à jour de 32 familles de 18 à 515 séquences.
- Participation aux *Cours RetNet de Bio-informatique* (Projet Européen RetNet)
- Formation Inserm en bio-informatique

CALENDRIER

2005	déploiement de PipeAlign sur une grille de calcul (GRID)
2005	génomes humain et murin d'Ensembl sur SGDB MySQL
2006	extension des services en ligne pour la génomique
2006	totalité des génomes complets d'Ensembl sur SGDB MySQL

LABORATOIRES IMPLIQUES

- Laboratoire de Bio-informatique et de Génomique Intégratives
- Plate-forme Bio-informatique de Strasbourg

UMR7104 - I.G.B.M.C.

MOTS CLES

haut débit; banques de données; extraction d'information; qualité de l'information; automatisation; technologie GRID; ressources bio-informatiques; services web;

PUBLICATIONS PRINCIPALES

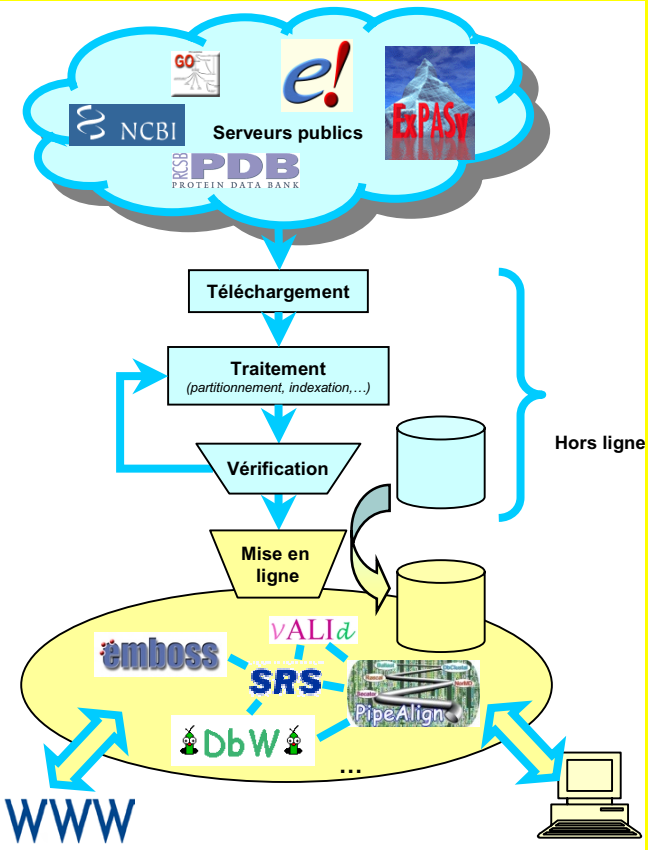
•[1] Bianchetti L.,Thompson J., Lecompte O., Plewniak F., Poch O. (2005) vALId : Validation of Protein Sequence Quality Based on Multiple Alignment Data. *Journal of Bioinformatics and Computational Biology* in press

•[2] Prigent V., Thierry J.C., Poch O., Plewniak F. (2005) DbW: automatic update of a functional family-specific multiple alignment. *Bioinformatics 21 :1437-42.*

•[3] Thompson J.D., Prigent V., Poch O. (2004) LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res. 32 :1298-307.

•[4] Plewniak F., Bianchetti L., Brelivet Y., Carles A., Chalmel F., Lecompte O., Mochel T., Moulinier L., Muller A., Muller J., Prigent V., Ripp R., Thierry J.C., Thompson J.D., Wicker N., Poch O. (2003) PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res. 31 :3829-32.*

SERVICES DE BIO-INFORMATIQUE



Les données téléchargées à partir des sites publics distributeurs des banques de données (NCBI, Expasy, PDB,…) sont d'abord traitées (partitionnement en sous-sections, indexation,…) pour une meilleure accessibilité. Le bon déroulement de ces opérations est ensuite vérifié manuellement avant de les mettre à disposition des utilisateurs par l'intermédiaire des logiciels intégrés par la plate-forme, soit directement par connexion sécurisée, soit sur le WWW.