Une application pour GDS ?

Antoine Vernois

Contexte

- Grands projets de séquencage de génomes complets (Homme, Souris, animaux,végétaux, unicellulaires)
- Croissance exponentielle de l'information biologique
  - Doublement de la taille des banques tous les 8-12 mois
  - 1-2 nouveaux génomes complets / mois
  - Plusieurs 10aines Go / mois
- Données brutes qui nécessitent une analyse

Objectifs des analyses

- Hypothèse sur la(les) fonction(s) d'une protéine inconnue
- Identifier des clusters de protéines
- Classer les protéines en famille
- Annoter les banques de séquences
- Croiser les informations biologiques

Ce que les biologistes attendent

- La puissance de calcul
- La capacité de stockage
- Un accès transparent
  - aux outils d'analyses
  - aux bases de données

Grid Protein Pattern Scanning

- ACI GRID 2002
- gridifier une application : PattInProt

Grid Protein Pattern Scanning

- ACI GRID 2002
- gridifier une application : PattInProt
- proposer des recommandations pour la gridification d'application bioinformatique

PattInProt : principe

- algo utilisé dans le cadre de l'ACI GriPPS
- principe
  - Un motif protéique plus ou moins complexe

    Cyclic nucleotide-binding domain signature 2

    [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]
  - Une séquence de protéine
  - Rechercher le motif dans la séquence
  - A l'identique ou en autorisant des erreurs

Un exemple

- cas d'utilisation
  - 1 motif vs 1 séquence
  - 1 motif vs banque(s) de séquences
  - banque(s) de motifs vs 1 séquence
  - banque(s) de motifs vs banque(s) de séquences

Les banques de données

- fichiers textes à plat

  ID CNMP_BINDING_2 ; PATTERN.

  AC PS00889 ;

  DT OCT-1993 (CREATED) ; OCT-1993 (DATA UPDATE) ; JUL-1998 (INFO UPDATE).

  DE Cyclic nucleotide-binding domain signature 2.

  PA [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
- de quelques Mos à plusieurs Gos
- mise à jour régulière
- accès principalement en lecture

Les banques de données (2)

- ex :
  - TrEMBL
    - séquences : 330 Mo
    - annotations : 1,6 Go
  - SwissProt
    - séquences : 60 Mo
    - annotations : 470 Mo

Algos bio : généralisons

- lectures séquentielles
  - une entrée = qq ko
- calcul souvent indépendant les uns des autres
  - petit mais nombreux
  - éventuellement synchro à la fin
- enchainement des traitements

Placement

- 1 calcul (pattern vs. protéine)
  - au pire qq sec. (après optimisation ;-)
- perte de temps lors des transfert
  - déplacer les calcul sur les bases
  - dupliquer les bases partout
  - recouvrement transfert/calcul

Placement

- 1 calcul (pattern vs. protéine)
  - au pire qq sec. (après optimisation ;-)

- perte de temps lors des transfert
  - déplacer les calcul sur les bases
    - engorgement
  - dupliquer les bases partout

  - recouvrement transfert/calcul

Placement

- 1 calcul (pattern vs. protéine)
  - au pire qq sec. (après optimisation ;-)

- perte de temps lors des transfert
  - déplacer les calcul sur les bases
    - engorgement
  - dupliquer les bases partout
    - pb de maj, de mises à jour
  - recouvrement transfert/calcul

Un résumé

- nombreux petits transferts
- nombreux petits calculs
- nombreux calculs avec +/- les mêmes entrées