Projet "GENOTO3D" de l'ACI "Masses de Données"

Compte-rendu de la réunion CASP6

jean-françois taly

CASP6

- **C**ritical **A**ssessment of Techniques for **P**rotein **S**tructure **P**rediction
- Juin-Aout 2005 : 6^{ième} session
- http://predictioncenter.llnl.gov/casp6/
- http://genome.jouy.inra.fr/~jftaly/

CASP6 : quelques chiffres

- 201 groupes d'experts
- 65 serveurs automatiques
- 64 cibles
- 90 domaines

4 catégories principales

- CM: comparative modeling
  - Homologie claire au niveau séquence
- FR-H: fold recognition homologous
  - Informations structurales
  - Support et cible homologues

4 catégories principales

- FR-A: fold recognition analogous
  - Support et cible structures similaires mais évolution convergente probable
- NF: new fold
  - Pas de support connu

CASP6 Vs CASP5

- Peu de modifications des méthodes
- Les meilleurs groupes sont les mêmes
  - Meta-serveurs meilleurs CM et FR-H
  - ROSETTA meilleur FR-A et NF

CASP6 Vs CASP5

- Confirmation profil-profil
- Meilleurs Alignements dans CASP6
- Amélioration grâce à augmentation PDB
- Meilleurs résultats sur cibles difficiles
- Mais moins de cibles difficiles

CASP6

- Cas facile :
  - Structure moyenne des supports est plus proche de la cible que la structure moyenne des modèles
- Cas difficile :
  - Structure moyenne des modèles est plus proche de la cible que la structure moyenne des supports

CASP6

- Meilleure stratègie
  - Alignement profil-profil séquence + struct2 + environnement
  - Définition des segments conservés et utilisation des supports correspondants
  - De-novo pour les boucles et segments orphelins
  - Assemblage avec contraintes issues prédiction de contacts + fonction d'énergie
  - Clustering : le + peuplé = natif

METHODES

- Comparaison séquences
- Threading
- Réseau neurones
- HMM
- Fonction énergie statistique
- Meta-serveurs

Comparative Modeling

- 42 des 90 domaines
- 24 CM/easy et 18 CM/hard
  - Easy si un blast suffit
  - Hard si 5 itérations de psiblast

Comparative Modelling

- Meilleurs groupes
  - Ginalski — *meta-serveur*
  - Venclovas et al. — *meta-serveur*
  - Kolinski et al. — *meta-serveur*
  - Skolnick et al. — *Threading + de novo*

Ginalski

- Meta serveur
  - 3D-jury
  - Meta-BASIC
  - ROSETTA + MAMMOTH

Ginalski : 3D-jury

- A partir des réponses de serveurs
  - Regroupe les modèles par similarité
  - Le groupe le plus peuplé = natif
- Problèmes
  - Si un seul serveur donne la bonne réponse
  - Pas d'amélioration du modèle
- Référence
  - (1)Bioinformatics Vol 19 (2003) 1015-1018

Ginalski : Meta-BASIC

- Fold recognition avec méta-profil
  - Profil de séquences
  - Prédiction de structures secondaires
- Référence
  - (2)Nucleic Acids Res., Vol 32 (2004) web issue

Ginalski : ROSETTA + MAMMOTH

- ROSETTA : Baker et al
  - De novo
- MAMMOTH : Ortiz et al
  - Comparaison structures
- CASP6
  - Détection de supports par comparaison entre modèles ROSETTA et protéines de structures connues

Skolnick : TASSER

- Détecte des fragments conservés avec méthode threading : PROSPECTOR-3
- Assemble ces fragments avec contraintes définies par threading : PHS parallel hyperbolic sampling
- SPICKER : sélection des modèles parmi decoys

FR-Homologues

- 23 des 90 domaines
  - Ginalski (431) *meta-serveur*
  - ...
  - Gene silico (311) *meta-serveur*
  - CBRC-3D (302) *meta profil-profil*
  - CHIMERA (275) *meta-serveur*
  - ROBETTA (228) *meta-serveur*

FR-Analogues

- 15 des 90 domaines :
  - ROSETTA (258) *de novo*
  - ...
  - GINALSKY (126) *meta-serveur*
  - SAM-T04 (121) *hmm + de novo*
  - JONES (112) *threading + réseau neurones + de novo*

KARPLUS : SAM-T04

- SAM-T2K : trouver séquences similaires
- Prédiction structure locale : réseau neurones + alignement multiple + 7 alphabets :
  - 4 en fonction des SSE
  - 2 en fonction des angles
  - 1 en fonction enfouissement

KARPLUS : SAM-T04

- SAM-T04 : two-tracks HMM, trouver des supports en tenant compte des prédictions précédentes.
  - http://www.soe.ucsc.edu/~karplus/casp6/
  - (6) SAM-T02 : proteins 53:491-496 (2003)
- Alignement cible-supports

KARPLUS : SAM-T04

- FRAGFINDER : librairie de fragments spécifique à la cible
- UNDERTAKER : algorithme génétique, chercher la meilleur conformation avec fragments. Favorise les états enfouis.

JONES-UCL : THREADER 3.5

- THREADER : threading profil-profile, paramètres alignement optimisés par algo génétique.
  - (7) FASEB vol 10:171-178(1996)

JONES-UCL : THREADER 3.5

- MgenTHREADER : réseau neurones, en entrée
  - Prédiction sse
  - Alignement des sse
  - Potentiel threading
  - Information sur la longueur
  - (8) Bioinformatics 19:874-881 (2003)

JONES-UCL : THREADER 3.5

- FRAGFOLD : assemblage de fragments de structures supersecondaires par recuit simulé
- MODCHECK : évaluation des modèles

New Fold

- 10 des 90 domaines
  - ROSETTA *de novo*
  - Kolinski *meta-serveur*
  - Ginalski *meta-serveur*
  - SAM-T04 *hmm*
  - Skolnick *threading*

BAKER : ROSETTA

- Alignement séquences
- Découpage en segments chevauchant de 9 résidus
- Profil-profil entre segments et banque de fragments
- Les fragments déterminent l'espace conformationnel

BAKER : ROSETTA

- Assemblage des segments
- Exploration espace conformationnel en minimisant une fonction d'énergie
- Recuit simulé en attribuant à la cible les coordonnées des fragments
- Clustering des modèles
- Le + peuplé = natif
- (10) J. Mol. Biol. 268:209-225

CONCLUSION

- Pas de grandes innovations
- Fonctions d'énergie actuelles : mauvaises structures avec énergie plus faible que native
- Utilisation de supports pour segments conservés
- Utilisation de librairies de fragments pour le reste

CASP6

MERCI DE VOTRE ATTENTION

Méthodes d'évaluation

- LCS : Longest Continuous Segments
- GDT-TS : Global Distance Test
- RMSd : Root Mean Square deviation
- AL0_P: Alignement with 0 shift
- AL4_P: with 1-4 shift tolerance

LCS
longuest continuous segment

- Chercher le segment le plus long pour lequel les paires de résidus sont situé à une distance inférieur à un cutoff (4Å)

GDT
Global Distance Test

- GDT : pourcentage de résidus pour distance < cutoff
- GDT_TS = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4,

GDT
Global Distance Test

- GDT : pourcentage de résidus pour distance < cutoff
- GDT_TS =

$$\frac{\text{GDT1Å} + \text{GDT2Å} + \text{GDT4Å} + \text{GDT8Å}}{4}$$

Sckolnick : PROSPECTOR-3

- Threading
  - Reference : proteins 56:502-518 (2004)
  - 3 fonctions de score :
    - Quasichemical
    - Protéines spécifiques, orientation indépendant
    - Protéines spécifiques, orientation dépendant
  - 2 profiles cible: séquence proche/éloignées
    - FASTA : 35% < id < 90%
    - FASTA : e-value < 10

Sckolnick : PROSPECTOR-3

- Profile supports : 3575 structures
  - >35 % id dans une famille
  - <35% id en dehors de la famille
- Algorithme itératif
  - 4 itérations
    1) Profile-profile de séquence > 10 structures
    2) Structures secondaires + prediction contacts
    3) Idem 2
    4) Idem2

Sckolnick : PHS

- Monte Carlo:parallel hyperbolic sampling
- Transformation du paysage énergétique
- Aplanir logarithmiquement les barrières d'énergies
- Reduir le temp passé dans un minima local
- (4) proteins 48:192-201 (2002)

Sckolnick : SPICKER

- Construire un grand nombre de modèles (MD)
- Clustering par paires en fonction RMSd
- Groupe le + peuplé = natif
- Superposition de toutes les structures d'un groupe sur la structure centrale
- Découpage en « sous chaînes » > 20 residus
- Moyenne de chaque sous chaîne
- (5) J. Comp. Chem. 25:865-871 (2004)