

Projet 1 : caractérisation de la famille des peroxydases de classe III

A partir de jeux de séquences de peroxydases de classe III et de classe I, faire un alignement multiple et essayer de construire HMM et/ou signatures sur les zones conservées.

Vous pourrez regarder cette publication, et peut-être télécharger les peroxydases de classe III mal annotées. [Automatic Prediction and Annotation: There Are Strong Biases for Multigenic Families - PMC \(nih.gov\)](#)

Fichier classIII : cp /mathe/save/BIS/PB_classIII.fasta .

Fichier classI: cp /mathe/save/BIS/PB_APx.fasta .

Fichier toutes la PeroxiBase : cp /mathe/save/BIS/peroxibase. Fa .

Obj : avoir un outil pour distinguer efficacement séquences de Classe I et Classe III. Evaluer votre méthode.

Projet 2 : THAP chez des organismes non animaux ?

- Interroger les banques par mots-clés
- Interroger les banques de domaines et chercher les THAP non animales
- Essayer avec BLAST, DeltaBLAST, Psi ou PhiBLAST ?
- Vérifier qu'elles ont bien (ou non) le domaine THAP
- Faire un alignement multiple des THAP d'éponge identifiées
- Comparer au pattern ou logo des bases de données de domaines

Obj : avoir trouvé toute la famille des protéines THAP chez des espèces non-animales ET AUSSI chez les métazoaires « basals ». Caractériser le domaine THAP chez ces organismes

Projet 3 : définition de la famille des RBOH

- Interroger les banques de domaines
- Trouver une protéine de la famille et regarder les liens vers les banques de domaines
- Constituer un ensemble de séquences appartenant à cette famille
- Faire un alignement de ces séquences
- Construire une signature, une HMM : pratt, hmmbuild
- Tester cette signature + HMM sur de nouvelles séquences

Obj : avoir identifié les résidus caractéristiques de cette famille, connaître la répartition taxonomique de cette famille de protéine, avoir établi une signature et une HMM spécifique

Projet 5 : caractérisation du domaine CBM1 des Oomycètes

- recherche des protéines à domaines CBM1 dans les banques : BLAST avec O42830, banques de domaines
- alignement multiple du domaine puis PRATT ou HMM
- MEME sur les séquences récupérées => comparaison du Logo avec celui des bases de données de domaines. On peut utiliser WebPRC

- comparaison avec Aphanomyces : récupérer les séquences qui ont le domaine IPR000254 (=CBM1) et tester le motif de MEME ou le pattern (ScanProsite) ou la HMM

Obj : avoir identifié les résidus caractéristiques du domaine CBM1 chez les Oomycètes. Montrer en quoi il diffère du CBM1 défini dans les bases de données de domaines, établi en majorité avec des séquences de champignons. Tester si le pattern ou la matrice (ou HMM) donne de bons résultats contre Aphanomyces (qui est aussi un Oomycète)

Projet 5 : protéines spécifiques du cerveau de plantes

Recherche par mots clés (et similarité ?) dans les banques.

Identifier des séquences : dans quelles banques ? chez quels organismes ?

Y a-t-il une différence entre les protéines 14-3-3 chez les plantes et les autres organismes ?

Y a-t-il une différence entre le domaine homeobox chez les plantes et les autres organismes ?

Obj : faire des hypothèses sur une annotation « étrange », comment la tester ? Qualité des informations dans les banques

Projet 6 et 7 : analyse d'un protéome

- Récupérer les séquences correspondant à PF00141 (=peroxidases) sur phytozome
- Tester BLAST ou BLAT sur Genotoul contre la Peroxibase
- Regarder quelques alignements
- Filtrer les résultats en définissant les critères : E-value, recouvrement, %id
- Ecrire un programme pour filtrer ou bien en ligne de commande (voir TP2 exercice 4)

**Obj : avoir utilisé BLAST et/ou BLAT en ligne de commande et avoir réussi à trouver des filtres
Combien de séquences validées et combien non validées ? Calcul de la Spécificité de ces données
Intérêt du travail sous Linux**

Projet 8 : annotation structurale

- Rechercher THAP9 zebrafish complète
- Modifier les paramètres du BLAST
- Utiliser CD-search (dans BLAST) sur partie en amont du match avec BX511023 traduite dans les 3 phases (cf publi réf Marchler-Bauer A et al.)
- Extraire la zone définie dans la publi et vérifier que c'est bien une protéine THAP
- La comparer par BLAST ou alignement multiple ou par 2 à d'autres protéines THAP, et constater que le domaine est incomplet. On peut aussi le vérifier avec ScanProsite
- chercher à partir de la séquence ADN d'autres sites d'épissage pour avoir le domaine complet => recomparer aux autres séquences ou banques de domaines

Obj : avoir retrouvé le résultat des auteurs. Montrer que leur domaine THAP est partiel. Si possible proposer un domaine complet