# TD 5 et 6 – Multiple alignment, Phylogeny and Psi-blast

## Première partie : psi-blast

Le mutant VTC2 d'*Arabidospis thaliana* est caractérisé par des niveaux très faibles en acide ascorbique (vitamine C). Le gène muté a été identifié comme étant le gène At4g26850 qui code pour la protéine référencée AAL07213.
On aimerait savoir de quel type de protéine il s'agit.

1) Recherchez la séquence protéique par Entrez. Avez-vous des informations sur la fonction moléculaire de cette protéine ?

Effectuez une recherche **psi-blast** (page blast du NCBI) à partir de la séquence d'*A. thaliana* dans la banque NR (avec en paramètre, un nombre maximum de séquences de 10 000).

2) Existe-t-il un domaine conservé d'après la recherche dans CDD ?

3) 1ère itération :
- Combien de hits obtenez-vous avec E<10 ?

*Remarque : par défaut, le nombre maximum de hits affichés (graphique et alignements) est 100. Pour avoir l'ensemble des hits, modifiez les options d'affichage en haut de page.*

- Dans quel(s) type(s) d'organismes ?
- Que pensez-vous de l'annotation des protéines homologues ?

4) 2ème itération
- Combien de hits obtenez-vous avec E<10 ?
- Dans quel(s) type(s) d'organismes ?
- Quel type de protéines apparaissent avec E<0.005 ?
- Quel motif ces nouvelles séquences possèdent-elles ?

5) 3ème itération
- Combien de hits obtenez-vous avec E<10 ?
- A quelle grande famille protéique appartient la séquence d'*A. thaliana* d'après les résultats obtenus ?
- D'après Interpro, il existe 3 sous-types à l'intérieur de cette superfamille. A quel sous-type appartient la séquence d'*A. thaliana* ?

Vous pouvez visualiser la place de VTC2 dans la voie métabolique de l'ascorbate en utilisant KEGG.

## Deuxième partie : Alignement multiple et Phylogénie

La plupart des organismes possèdent une lysyl-ARNt synthétase de classe II, cependant quelques bactéries et archées possèdent une lysyl-ARNt de classe I. Des séquences représentatives de ces dernières sont disponibles : **/home/lecompte/lysRS.seq**

**1) Alignment construction using clustalx**
Connectez-vous au serveur titus :
**`ssh -X depulp1@titus.u-strasbg.fr`**

Créez votre répertoire de travail (**mkdir**). Copiez l'alignement dans votre répertoire (**cp**).
Lancez le program **clustalx**, chargez les séquences et alignez-les (**do complete alignment**).
Sauvez l'alignement au format msf.

**2) Alignment visualization using seqlab.**
Initialisez la suite de programme GCG (**gcg**).
Lancez l'éditeur d'alignements multiples seqlab (**seqlab&**).
Chargez l'alignement dans seqlab en mode editeur.

- L'alignement vous semble-t-il correct ?
- La conservation est-elle homogène pour l'ensemble de la protéine ?
- Identifiez un motif fortement conservé.
- Quels sous-groupes distinguez-vous ?

**3) Construction of phylogenetic trees with phylowin.**
Il nous faut un fichier au format fasta. Changez le format du fichier avec la commande seqret.

**seqret fichiermsf**

Ouvrez le fichier au format fasta avec le programme phylowin :

**phylowin fichierfasta &**

Sélectionnez toutes les espèces et tous les sites.

a) Construisez l'arbre selon la méthode du maximum de parcimonie

- Combien de sites sont informatifs ?
- L'arbre est présenté sous la forme enracinée. Etes-vous d'accord avec la position de la racine ?
- La topologie est-elle en accord avec vos observations sur l'alignement ?

b) Construisez l'arbre selon la méthode du neighbor-joining avec les options :
*global gap removal* (valeur par défaut), *observed divergence*, *bootstrap*=100

- Combien de sites sont utilisés ?
- La racine est-elle bien positionnée ?
- Commentez les valeurs de bootstrap.

c) Refaire l'arbre selon la méthode du neighbor-joining avec l'option *pairwise gap removal*, *observed divergence*, *bootstrap*=100

- Combien de sites sont utilisés ?
- Comparez l'arbre obtenu avec l'arbre précédent. Comment pouvez-vous expliquer les différences ?

## Troisième partie : Etude d'une famille

### Recherche d'une séquence (http://srs6.ebi.ac.uk/)

Rechercher la protéine mRNA guanylyltransferase de *Saccharomyces cerevisiae* dans la banque Uniprot.

- Quelle est la longueur de la protéine ? sa localisation cellulaire ? sa fonction ?
- Quel est le site actif de la protéine ? Notez le résidu catalytique et son environnement immédiat afin de pouvoir le retrouver dans l'alignement.
- Récupérez la séquence au format fasta.

### Alignement de la séquence MCE1_YEAST avec ses séquences homologues

On utilisera la suite de programmes PipeAlign :
http://bips.u-strasbg.fr/PipeAlign/.

Soumettre la séquence MCE1_YEAST au format Fasta et notez votre numéro de session.

### Etude de la famille de protéines

**1. Conservations (choisissez l'alignement DbClustal)**

Observez les positions de forte conservation dans l'alignement.

- L'alignement obtenu est-il globalement correct selon vous ?

*Vous pouvez visualiser le profil de conservation pour voir les ancres mises en évidence par Ballast.*

- Est-ce que toutes les séquences font bien partie de la famille ? Sont-elles toutes correctement alignées ?
- Quelles séquences ont été éliminées par Léon ? Etes-vous d'accord ?

**2. Erreurs de séquences et d'alignement**

Existe-t-il des erreurs de séquences dans l'alignement ?

**3. Organisation en domaines**

L'organisation en domaines vous paraît-elle homogène au sein de cette famille ?

**4. Sous-familles**

- Sur la base de l'organisation en domaines et des conservations, quels groupes distinguez-vous ?
- Quelles sont les différences fonctionnelles entre ces groupes d'après les annotations ?
- Est-ce que ces groupes sont en accord avec les groupes taxonomiques ?

## First part: psi-blast

The VTC2 mutant of *Arabidopsis thaliana* is characterized as showing very low ascorbic acid (vitamin C) levels. The mutated gene is identified as the gene At4g26850 encoding the protein with access AAL07213 in the Refseq database.
We would like to obtain further information on this protein.

1) Search for the protein sequence using Entrez.
- Do you have information about the molecular function of this protein?
Perform a psi-blast search (NCBI) with the *A. thaliana* sequence in the NR database (default parameters with a maximum of 10 000 sequences).

2) Does the protein exhibit a conserved domain according to the CDD search?

3) First iteration:
- How many hits do you obtain with E<10?
*By default, the maximum number of hits (graphical overview and alignments) is limited to 100. Reformat the results to obtain all the results.*
- In which type(s) of organisms?
- What do you think about the annotations of homologous proteins?

4) Second iteration:
- How many hits do you obtain with E<10?
- In which type(s) of organisms?
- Which type of proteins appears with E<0.005?
- Which pattern do these proteins exhibit?

5) Third iteration:
- How many hits do you obtain with E<10?
- To what large family of proteins does the sequence of *A. thaliana* belong according to these results?
- According to Interpro, there are 3 subfamilies within this large family. What is the subfamily of the *A. thaliana* sequence?

You can visualize the VTC2 gene in the ascorbate metabolic pathway using KEGG website.

## Second part: multiple alignment and phylogeny

Most organisms exhibit a lysyl-tRNA synthetase belonging to the class II of synthetases but some bacteria and archaea have a lysyl-tRNA synthetase belonging to the class I. Some representatives of the latter sequences are available: **/home/lecompte/lysRS.seq**

**1) Alignment construction using clustalx**
We will use the server titus:
**ssh -X depulp1@titus.u-strasbg.fr**

Create your directory (**mkdir**). Copy (**cp**) the alignment in your directory.
Launch **clustalx**, load sequences in the editor then do complete alignment.
Save the alignment in msf format.

**2) Alignment visualization using seqlab.**
Launch the GCG suite of programs (**gcg**) and the seqlab editor (**seqlab&**).
Load the alignment in seqlab.

- Does the alignment seem correct?
- Is the conservation homogeneous along the whole protein?
- Identify a strongly conserved pattern.
- Which subgroups can you distinguish?

**3) Construction of phylogenetic trees with phylowin.**
We need a file in fasta format. Change the alignment format using the seqret command:

**seqret fichiermsf**

Launch phylowin:

**phylowin fichierfasta &**

Select all the species and all the sites.

a) Construct the phylogenetic tree with the *maximum parsimony* method
- How many sites are informative?
- The tree is shown as a rooted tree. Do you agree with the root position?
- Is the topology in agreement with your observations on the multiple alignment?

b) Construct the phylogenetic tree with the *neigbor-joining* method (*global* gap removal, observed divergence, bootstrap=100).
- How many sites are used?
- Is the root correctly positioned?
- Comment on the bootstrap values.

c) Construct the phylogenetic tree with the *neigbor-joining* method (*pairwise* gap removal, observed divergence, bootstrap=100).
- How many sites are used?
- Compare the tree with the previous one. How can you explain the differences?

## Third part: family study

### Sequence search (http://srs6.ebi.ac.uk/ )

Search for the protein sequence of the mRNA guanylyltransferase of *Saccharomyces cerevisiae* in the Uniprot database.

- What is the sequence length? Its cellular localisation? Its fonction?
- Where is the protein active site? Note the catalytic residue and the adjacent residues to localize the active site in the multiple alignment.
- Copy the sequence in fasta format.

### Alignment of the MCE1_YEAST sequence with its homologs

We will use the PipeAlign suite of programs:

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

Submit your sequence in fasta format and note your session ID.

### Analysis of the protein family

**1. Conservations (Choose the DbClustal alignment)**

Observe highly conserved positions in the alignment.

- According to you, is the alignment globally correct?

*You can visualize the Ballast conservation profile in a new window to see the ballast anchors.*

- Do all sequences belong to the family? Are they all correctly aligned?
- Which sequences have been eliminated by Leon? Do you agree?

**2. Sequence errors**

Can you detect some sequence errors?

**3. Domain organization**

Is the domain organization homogeneous within the family?

**4. Subfamilies**

- According to domain organization and conservations, which groups can be distinguished?
- What are the functional differences among these groups according to the annotations?
- Are these groups in agreement with the taxonomy?