TP GMIN206 - Outils Bioinformatique

V. Berry – temps nécessaire = 6h

2 avril 2013

Résumé

Outils bioinformatiques couverts :

- Site NCBI Taxonomy Browser
- Site Tree of Life Web
- Sites iSpecies, wikiSpecies
- Sites Orthomam, PhyloExplorer
- Site CompPhy pour la construction de super-arbres et la comparaison d'arbres
- Logiciel MEGA : interface aux programmes d'alignement et de construction de phylogénies
- Interface de MEGA au site Entrez et au logiciel BLAST

NOTE : les différents fichiers de données mentionnés peuvent être trouvés à l'adresse `www2.lirmm.fr/~vberry/COURS/GMIN206/MEGA`.

Première partie

Sites pour la taxonomy + introduction à MEGA (3h)

---

1 Données réelles : phylogénie des mammifères placentaires

Nous nous intéressons ici plus particulièrement aux rongeurs, des animaux compris dans le groupe des mammifères placentaires (*Eutheria*), pour lequel certaines relations phylogénétiques sont encore incertaines.

Pour étudier les relations phylogénétiques chez les rongeurs, nous allons d'abord utiliser l'exon 28 du gène codant le facteur de von Willebrand, séquencé chez de nombreux taxons pour tester l'hypothèse de la monophylie des rongeurs : est-ce que ces animaux constituent un seul groupe phylogénétique ou non.

Question 1 *Téléchargez le fichier* `vwf-12.nuc` *depuis la page web donnée au début de ce TD.*

Un bref regard au fichier en question nous apprend que le jeu de données inclut 12 espèces parmi lesquelles l'incontournable Homo Sapiens.

Question 2 *Pour connaître la frimousse des charmantes espèces que nous allons étudier demandez une photo de chacune grâce aux sites* WikiSpecies *et* iSpecies.

Question 3 *Nous étudions ici principalement la phylogénie des rongeurs, pourquoi avoir inclus des espèces n'appartenant pas à ce groupe taxonomique ?*

1.1 Etat des relations taxonomiques connues

Un certain nombre de sites recensent les relations taxonomiques entre les espèces connues. Ces relations sont découpés en différents niveaux comme espèce, ordre, genre, classe, ...

Parmi ces sites, le *NCBI Taxonomy Browser* fait référence (mais voir aussi le site du *GBIF*).

Question 4 *Utilisez le site du NCBI pour identifier parmi ces espèces, lesquelles ne sont pas des rongeurs.*

Question 5 *Quel est le nom du plus petit groupe taxonomique qui contient toutes les espèces du jeu de données ? De quel type de groupe s'agit-il (genre, ordre, classe, ...) ?*

Question 6 *Utilisez ce site pour identifier à quel* ordre *appartient le lièvre ; sous quel nom est connu le groupe taxonomique regroupant les rongeurs et l'ordre auquel appartient le lièvre ?*

Question 7 *A quel* super-ordre *appartient le hérisson ; même question pour le daman.*

Plusieurs sites internationaux recensent les relations montrées scientifiquement entre les espèces connues. Le site du *Tree of Life* (`http://tolweb.org/tree/`) est l'un de ces sites bien connus.

Question 8 *En parcourant ce site, établissez une réplique de l'arbre phylogénétique connu pour les 8 rongeurs de notre étude. De quand date l'étude sur laquelle est basée la phylogénie représentée ? Quelles sont les clades non résolus ?*

1.2 Premières analyses

Nous allons faire la plupart des manips depuis le logiciel **MEGA** que vous trouverez en salle de TP sur une session windows. Localisez ce programme dans le menu *Démarrer*.

Pour utiliser le jeu de données dans MEGA il faut d'abord le convertir dans un format qu'il peut lire.

Question 9 *Démarrez MEGA et chargez le jeu de données vwf-12.nuc en demandant à le convertir depuis le format actuel (Phylip, séquences entrelacées, c-a-d* interleaved*).*

A chaque fois que MEGA charge un fichier ou effectue une analyse, il représente l'objet en question sous la forme d'une petite icône carrée. Vous pouvez sauvegarder votre session (et toutes ses icônes) en cliquant sur l'icône *Data* du menu graphique, puis en choisissant *Save data session to file ....*

Question 10 *Les données chargées se présentent-elles en vrac ou sous la forme d'un alignement ? Quel indice vous permet de répondre ?*

Passons maintenant à l'inférence d'une phylogénie depuis ces données :

Question 11 *Lancez dans un premier temps une construction de phylogénie par parcimonie depuis les données chargées. En demandant que la phylogénie soit construite sur la base de l'heuristique CNI (close neighbor interchange). Expliquez comment fonctionne cette heuristique.*

Question 12 *Dans la visualisation obtenue, en utilisant les icônes appropriées demandez à avoir une représentation courbe de la phylogénie, et positionnez la racine correctement, c-a-d sur la branche menant à l'outgroup le plus éloigné des rongeurs (cf votre étude depuis TolWeb).*

Vous remarquerez que dans la visu de la phylogénie construite, que pour un caractère choisi (initialement celui de la colonne 1), on peut avoir la reconstruction des états ancestraux pour ce caractère.

Question 13 *Utilisez la boîte à outil pour épaissir les branches à 2pts et représentez un carré rouge devant les 4 espèces outgroup du jeu de données. Que remarquez-vous sur la position de ces 4 outgroups ? Est-ce que c'est ce à quoi vous vous attendiez ?*

1.3 Analyses complémentaires

Comme raison de cette situation étonnante, plusieurs explications peuvent être invoquées qu'il nous faut étudier. Commençons par mettre en doute la dernière étape de l'analyse : nous avons utilisé une méthode heuristique (approchée) pour obtenir la phylogénie en parcimonie.

Question 14 *Relancez une analyse de parcimonie en utilisant cette fois la méthode Maxi-Mini Branch and bound. Il s'agit d'une méthode exacte et donc bien plus coûteuse en temps calcul, le problème informatique en question étant* NP-complet [1]. *Le résultat est-il différent de ce que vous aviez obtenu en première analyse ?*

En revenant sur la démarche des analyses, un autre soucis est peut-être le type de méthode employé pour construire la phylogénie : la parcimonie n'est peut-être pas le type de méthode le plus adapté ? Pour tester cette hypothèse, nous allons reconstruire une phylogénie depuis une méthode de distances.

Il est reconnu que la méthode de parcimonie est d'autant plus efficace que les taux d'évolutions sont faibles et que les taux d'évolutions sont constants sur l'ensemble des branches de la phylogénie.

Question 15 *Calculez la distance évolutive moyenne entre les séquences du jeu de données. Qu'en déduisez-vous ?*

Les auteurs de la méthode de distances appelée *Neighbor Joining* affirment (de manière peut-être un peu optimiste) que leur méthode peut être utilisée jusqu'à une distance moyenne de 1.0 entre séquences (quand la distance est évaluée au sens du modèle de Jukes et Cantor.

Question 16 *Effectuez le calcul au sens de ce modèle, pour voir si nous pouvons utiliser leur méthode en remplacement de la parcimonie. Quelle conclusion ?*

Question 17 *Effectuez maintenant une analyse par la méthode de* neighbor joining *(plutôt que Minimum Evolution dans un cas de distances relativement importantes entre séquences). Utilisez pour convertir les séquences en distances le modèle Kimura à deux paramètres. Quelle conclusion pour la phylogénie obtenue depuis cette méthode ?*

Question 18 *Au passage, est-ce que les branches de cette phylogénie ont des longueurs plutôt semblables ou non ? (autrement dit est-ce que la parcimonie était appropriée à l'analyse de ce jeu de données ?)*

1. wikipedia

Si le résultat de l'analyse de reconstruction phylogénétique est encore étonnant, le modèle choisi pour inférer les distances entre séquences (et estimer les substitutions silencieuses) est peut-être en cause...

Question 19 *Pour tester cela relancez une analyse de Neighbor Joining en utilisant cette fois un modèle de vraisemblance avec loi Gamma pour estimer les distances. Quelle conclusion ?*

1.4 Retour sur le jeu de données

Malgré des analyses différentes, le problème de non-monophylie des rongeurs persiste, il nous faut donc revenir vers les séquences elle-mêmes.

La séquence de l'homme est particulièrement distante des autres séquences. Revenez sur l'alignement et vérifiez-le en regardant les séquences ADN.

Question 20 *Cela apparait-il aussi quand on examine les séquences du point de vue des acides aminés représentés par les codons ?*

Par défaut la transformation nucléotide → acides aminés s'opère dès la première colonne. Cela peut être un soucis si le cadre de lecture ne commence en fait pas à cette position mais à la suivant ou celle d'encore après (vous savez que la variabilité permise par le code génétique n'est pas la même sur chacune des 3 positions du codon).

Question 21 *Pouvez-vous trouver comment faire pour que MEGA vous donne la traduction en acide aminés en partant du 2ème (ou du 3ème) site et non du 1er site de l'alignement actuel ?*

A ce propos, ce que l'on fait souvent quand le signal est trop saturé (trop de quantité évolutive entre les séquences) c'est de construire une phylogénie uniquement depuis les 1ères et 2èmes positions du codon.

Question 22 *Est-ce possible dans MEGA, et si oui, de quelle façon ? Est-ce que cela règle les soucis rencontrés précédemment ?*

Pour revenir à la séquence de l'homme dans le jeu de données, celle-ci est probablement erronée (un exon différent de celui sélectionné chez les autres espèces a peut-être été choisi par mégarde).

Pour étudier l'influence de cette séquence sur l'ensemble de l'analyse nous allons la désactiver et reconduire deux analyses rapidement :

Question 23 *Dans l'icône* ***TA..*** *correspondant au jeu de données, décochez la séquence de l'homme, puis redemandez à MEGA la distance moyenne entre séquences (restantes). Quelle conclusion ?*

Question 24 *Pour voir si cette séquence était déstabilisante, relancez une analyse de parcimonie (adaptée à des taux de 0.25) et une analyse de Neighbor Joining maintenant que cette séquence est désactivée. Est-ce que certains soucis sont corrigés ?*

Deuxième partie

Plus loin avec Mega

1.5 Acquisition d'une nouvelle séquence

Dans un autre module, vous avez appris comment vous servir de SRS et Entrez pour collecter des données depuis les banques de données internationales.

Question 25 *Mettez en pratique vos connaissances pour chercher dans ces banques la séquence de l'exon 28 du gène VWF.*

Enregistrez la séquence correspondante dans un fichier fasta du nom de `Homme-vwf28.fa` par exemple.

MEGA permet sans quitter le logiciel d'aller chercher des séquences pour compléter un jeu de données.

Question 26 *Depuis l'icône graphique* Alignement*, demandez à effectuer une recherche de séquence sur le web (option* Query Databanks *ou* BLAST search*). Quand vous avez trouvé la séquence qui vous intéresse, cliquez sur l'icône "+" en haut du navigateur de MEGA, la séquence choisie est automatiquement ajoutée aux séquences du jeu de données*

Pour la suite des questions, vous dé-sélectionnerez l'ancienne séquence pour l'homme et utiliserez à la place la nouvelle séquence récupérée par le web.

Question 27 *Relancez un alignement en choisissant le programme Clustal. Faîtes un autre alignement en utilisant le programme Muscle. Comment se comparent ces alignements ?*

Depuis l'alignement qui vous semble le plus correct, relancez une analyse phylogénétique par méthode de parcimonie ainsi qu'une analyse par méthode Neighbor Joining.

Question 28 *Quels changements obtenez-vous pour ces phylogénies ? Est-ce que l'homme rejoint la place à laquelle il était attendu phylogénétiquement d'après les site du* Tree of Life *et celui de la taxonomie du* NCBI *?*

1.6 Artefact de reconstruction - bootstrap

Question 29 *Dipodomys (le rat-kangourou) est un rongeur lié à la phylogénie par une longue branche. A la lumière de ce fait, pouvez-vous ré-interpreter le placement de ce taxon – commun aux phylogénies construites précédemment ?*

Pour mieux mesurer ce phénomène et mesurer la force des différences entre phylogénies précédentes, nous allons conduire des analyses bootstrap afin d'obtenir une mesure de confiance pour chaque clades dans le cas des deux méthodes précédentes d'inférence. Relancez une analyse de parcimonie en demandant un test bootstrap (choisissez 500 réplicats bootstrap).

Question 30 *Notez la présence de deux arbres dans la fenêtre de résultats (deux onglets). A quoi correspondent-ils ?*

Question 31 *Est-ce que les valeurs de confiance obtenues sont importantes ? Quels clades conserveriez vous à* 95% *de confiance ? Montrez-le en réduisant à ces clades la phylogénie à l'aide des options de la fenêtre de visualisation.*

Question 32 *Pour revenir à la présence de longues branches dans le jeu de données, pouvez-vous expliquer (en parcimonie au moins) pourquoi le hérisson et l'écureuil sont groupés ensemble avec une forte valeur de support statistique bootstrap ?*

Le bootstrap peut être une autre manière de montrer à quel point la séquence erronée de l'homme initialement présente dans le jeu de données déstabilisait (ou pas) les analyses.

Question 33 *Comparez les valeurs bootstrap obtenues depuis le jeu de données initial et celui où la séquence de l'homme a été remplacées sur une même méthode de reconstruction phylogénétique.*

1.7 Influence des outgroups

Il est connu que si les outgroups utilisés sont trop distants du groupe que l'on étudie (ici les rongeurs) ou trop peu nombreux, alors ils peuvent venir s'enraciner n'importe où. Pour

étudier la sensibilité des arbres produits, nous allons compléter le jeu de données en ajoutant progressivement des membres supplémentaires des groupes Afrothériens et Laurasiathériens. A notre disposition nous avons :

Laurasiathér.	Manis (pangolin)	Felis (chat)	Equus (cheval)	Physeter (cachalot)
num accés séq	U97535	U31613	U31610	AF108834
Afrothér.	Orycteropus (Aadvark)	Dugong	Loxondonta (éléphant)	Tenrec (tangue)
num accés séq	U31617	U31608	U31615	AF390536

Question 34 *Choisissez une espèce parmi celles indiquées dans le tableau précédent (ou d'autres) et obtenez sa séquence pour l'exon 28 du gène VWF par l'intermédiaire du programme MEGA. Calculez un nouvel alignement pour tenir compte de cette séquence.*

Enregistrez le nouveau jeu de données sous le nom `vwf-13` dans votre répertoire de travail.

Question 35 *Calculez un nouvel alignement pour tenir compte de cette séquence.*

Remarque : il est possible dans l'alignement par Muscle piloté par MEGA d'indiquer des options de ligne de commande. Pouvez-vous demander ainsi à réaliser un alignement profil, c-a-d ne revenant pas sur l'alignement des séquences déjà présentes et alignant juste la nouvelle séquence à cet alignement précédent ?

Question 36 *Regardez l'alignement obtenu :*

1. *comment sont répartis les gaps*
2. *que pensez-vous du début de l'alignement ? Quelle raison/manipulation a pu conduire à cette situation ?*
3. *que pensez-vous de l'alignement dans son ensemble ?*

Question 37 *Construisez une phylogénie par méthode de distances sur le jeu à 13 taxons avec les mêmes programmes que précédemment. Que constate-ton ?*

Un doute vous assaille soudain : est-ce que les résultats obtenus ne seraient pas partiellement satisfaisant uniquement parce que vous auriez choisi/obtenu une mauvaise séquence ?

Question 38 *Pour lever ce doute, demandons à obtenir la même séquence par un autre site web et une autre banque de données aussi si possible. Pour celà, allez sur le site SRS de l'EBI (European Bioinformatic Institute) et recherchez la même séquence que celle obtenue précédemment, mais cette fois en effectuant la recherche dans la base EMBL (qui recoupe fréquemment ses entrées avec GenBank mais qui est indépendante).*

1.8 Méthode de Maximum de vraisemblance

Parce qu'elle est plus robuste, nous pouvons espérer que la méthode du maximum de vraisemblance ne fera pas les erreurs faites par la parcimonie et les méthodes de distance.

Question 39 *Utilisez dans un premier temps MEGA pour déterminer le modèle d'évolution des séquences le plus approprié aux données :* Find Best Model. *Quel est le meilleur modèle au sens du critère BIC ?*

Lancez maintenant une reconstruction de phylogénie avec une méthode de maximum de vraisemblance se reposant sur le modèle d'évolution des séquences le plus adéquat, comme déterminé ci-dessus. Demandez aussi un processus bootstrap avec 500 réplicats (ou 200 si trop long).

Question 40 *Est-ce que les valeurs bootstrap sont meilleures que pour les phylogénies construites précédemment ? Quelle(s) explication(s) ?*

Question 41 *Est-ce que le problème de longue branche qui affectait précédemment la phylogénie est toujours présent ?*

1.9 D'autres séquences encore

En dernier recours avant de jeter le gène VWF à la poubelle, nous allons essayer de le compléter par plus d'espèces du groupe des hérissons, afin d'essayer de casser la longue branche qui le relie au reste des espèces étudiées.

Question 42 *Ajoutez cette fois 5 nouvelles séquences provenant d'espèces s'étalonnant entre le hérisson et les autres espèces déjà étudiées. Pour localiser ces nouvelles espèces, vous pouvez vous servir du tableau donné plus haut dans l'énoncé, ou vous rendre sur le site du* Taxonomy browser*, y chercher le Hérisson et sélectionner le nom de quelques espèces voisines et pas trop voisines. Ensuite dans l'interface de MEGA à* Entrez *allez chercher les séquences pour ces espèces(attention à bien sélectionner l'exon 28 de VWF, sinon vous n'aligneriez des séquences qui n'ont rien à voir).*

Question 43 *Passez par l'étape d'alignement et ensuite de construction d'une phylogénie (utilisez par exemple le maximum de vraisemblance). Est-ce que les outgroup ont maintenant une place normale, c-a-d les rongeurs sont-ils regroupés ? Si oui, quelle est la phylogénie suggérée pour les rongeurs, exportez-la dans un fichier* `vwf-X.nwk` *au format newick (ici* `X` *est le nombre de séquences de votre jeu).*

1.10 Arbres de gènes / arbre d'espèces ?

Il est possible que certains événements conduisent à proposer une histoire du gène vwf qui n'est pas celle de la phylogénie : duplication et pertes de gènes, gènes paralogues, transferts horizontaux.

Toutefois, dans le cas présent, les experts pensent que le gène ne contient pas assez de sites pour traduire clairement les clades sous-jacents de la phylogénie des espèces.

Question 44 *Quel procédé pouvez-vous imaginer pour approuver ou réfuter cet avis ? Procédez à l'expérience. Quelle conclusion ?*

1.11 Signal présent dans d'autres gènes – Construction d'un super-arbre

Pour nous faire une autre idée de la phylogénie des rongeurs, nous pouvons aussi avoir recours à d'autres gènes. Nous utiliserons ici les gènes `a2ab` et `irb` pour lesquels nous avons un alignement sur de nombreuses espèces incluant les rongeurs.

Question 45 *Analysez les fichiers* `irb.nuc` *et* `a2ab.nuc` *avec MEGA. Quelle conclusion pour les rongeurs dans l'ensemble ?*

Il est difficile de comparer à l'oeil tous les arbres obtenus et impossible de le faire par une méthode de consensus classique car ces gènes n'ont pas le même ensemble de taxa.

Nous allons donc avoir recours à une méthode de *superarbre* pour obtenir un consensus de ces arbres. Nous allons utiliser la méthode nommée *physic_ist*. Pour cela, nous allons passer par la plateforme *CompPhy* développée au LIRMM. Rendez-vous à l'adresse `http://www.atgc-montpellier.fr/compphy/` créez chacun un compte puis groupez-vous par deux pour créer un projet (chacun sur sa machine) : l'un des deux crée le projet en invitant son binôme à participer.

Question 46 *Dans MEGA, enracinez chaque arbre de gène que vous avez obtenu au bon endroit puis exportez-le dans un fichier au format newick. Ensuite, collectez par copier/coller les formes newick correctement enracinées dans un même fichier que vous nommerez* rodents.nwk.

Question 47 *Chargez maintenant la collection d'arbres dans votre projet (attention, seul celui qui a la main sur le projet (voir cadre en haut à droite) peut mettre à jour celui-ci).*

Question 48 *Une fois que la collection de trois arbres est chargée lancez et après avoir vérifié qu'ils sont chacun enraciné au bon endroit (sinon passez par l'outil d'enracinement), lancez le calcul d'un super-arbre par la méthode Physic_IST (utilisez les paramètres par défaut). Cette méthode combine les arbres de gènes en un superarbre résumant cette collection initiale.*

Question 49 *Pouvez-vous ainsi établir les confits entre arbres de gènes et les clades sur lesquels la collection est majoritairement d'accord ? Pour souligner graphiquement les points de désaccord, vous pourrez utiliser l'outil MAST.*

Question 50 *Eventuellement, pour obtenir un super-arbre plus résolu, essayez plusieurs variantes autour de ces seuils (bootstrap à 0.7 par exemple, et STC de 0.8 à 1).*

Afin de clarifier l'histoire des rongeurs, nous pouvons fournir plus d'arbres de gènes à la méthode de superarbre : plus on dispose d'arbres de gènes plus on s'extrait des soucis propres à chacun (transferts éventuels, paralogie si de mauvaises séquences sont collectées, etc). D'après un éminent spécialiste du groupe des rongeurs, on peut faire confiance aux gènes suivants :

- GHR : Growth Hormone Receptor
- BRCA1 : BReast CAncer 1
- CNR1 : cannabinoid receptor 1
- RAG1 et RAG2 : recombination activating gene 1 et 2

Question 51 *Collectez les arbres connus pour ces gènes dans les banques de séquences (par exemple* ENSEMBL *ou le site Montpellierain dédié aux mammifères :* Orthomam*). Une fois obtenu de nouveaux arbres de gènes et après les avoir enracinés correctement, ajoutez les arbres à votre projet, puis relancez la méthode physic_ist pour voir vers quelle phlyogénie des rongeurs converge le superarbre.*

1.12 Conclusion

Après tout ce temps passé à courir derrière les rongeurs, vous vous demandez si vous avez obtenu une résolution phylogénétique en accord avec ce qui est connu pour les rongeurs depuis 2005 (date de la phylogénie stockée sur le site Tree Of Life (*tolweb*). Certaines avancées sur ce groupe sont implémentées dans la taxonomie du NCBI.

Pour connaître cette taxonomie, nous pouvons explorer manuellement le site du *Taxonomy Browser* vu au début de ce TP, en localisant successivement toutes les espèces que nous avons étudiées aujourd'hui.

Question 52 *Essayez de tracer sur papier la phylogénie des rongeurs en partant de 4 à 5 espèces étudiées aujourd'hui. Seriez-vous capable de générer le codage newick de cette phylogénie ?*

Une deuxième solution pour obtenir plus rapidement la vision acceptée actuellement sur les rongeurs que nous avons étudié consiste à utiliser le portail de gestion d'arbres mis en place à Montpellier : `http://www.ncbi.orthomam.univ-montp2.fr/phyloexplorer/`.

Question 53 *Entrez votre collection d'arbres rodents.nwk sur ce site et visualisez la phylogénie connue sur le site du NCBI pour ces organismes. Attention, il vous faudra utiliser les noms connus pour ces espèces (noms scientifiques ou noms communs en anglais), que vous pouvez obtenir sur le site du NCBI ou sur le GBIF. Vérifiez que la phylogénie NCBI montrée par PhyloExplorer correspond à ce que vous venez de tracer à la main, et à ce que vous avez inféré précédemment dans l'analyse de super-arbres*

Bravo, vous avez fait progressé la science sur la résolution des relations phylogénétiques entre rongeurs. La phylogénie obtenue peut être utilisée à différentes fins : sélectionner les espèces à protéger, choisir une espèce pour effectuer des tests dans la lutte contre certaines maladies, estimer les vitesses de diversifications récentes, ...

Question 54 *Dans le portail* CompPhy*, on peut avoir des soucis à l'entrée des arbres dans un projet car le format des fichiers n'est pas vérifié avant de créer les imagettes qui leur correspondent. Pouvez-vous créer un script simple qui parcours chaque fichier d'arbres soumis au portail et indique si un arbre au format Newick est mal formé, et à quelle position ? Attention :*

- *le contenu d'un même arbre peut être étalé sur plusieurs lignes*
- *certains arbres contiennent à la fois des longueurs de branche et des valeurs de support.*
- *le format des longueurs de branches peut comporter la lettre 'E' (ou 'e') ainsi que les symboles '+' et '-'.*
- *on demande de transformer en symbole '_' tout caractère non alphanumérique*
- *est-il possible d'accepter qu'un nom de taxon commence par un numéro ? (voire ne soit composé que de numéros et de symboles) ? (voir manuellement ce que supporte l'outil de dessin* `www.scriptree.org` *utilisé dans* ComPhy*.*

Troisième partie

Trouver les séquences homologues à une séquence

Supposons maintenant que dans le cadre d'un travail sur *Thermogata petrophilia*, vous veniez de collecter par des manipulations expérimentales (PCR, ...) une séquence, que vous avez identifié comme étant la séquence dont l'accession number est *CP000702*.

Utilisons MEGA pour constituer un jeu de données composé des séquences orthologues à votre séquence.

Question 55 *Pour cela, utilisez dans un premier temps l'option* `Align` *puis* `BLAST Search` *et indiquez CP000702 dans le champ "*Enter accession number, gi, or FASTA sequence*". Comme* `Database` *indiquez* Nucleotide collection (nr/nt) *puis choisissez l'outil* blastn *(somewhat similar sequences).*

Pour déterminer quelles séquences sont orthologues, et non paralogues ou xénologues à notre séquence initiale, ou simplement homologues par chance (partage d'un seul domaine), il n'y a pas de règle bien établie.

Question 56 *Sélectionnez les séquences qui qui couvrent (*coverage*)* $\geq 60\%$ *de notre séquence initiale* ***et*** *qui ont une* E-value $< 10^{-3}$*. Pour aller voir les séquences qui nous intéressent, nous allons cliquer sur leur valeur dans la colonne* Max score*, ce lien nous amène à l'alignement entre la séquence et notre séquence requête.*

Pour récupérer les séquences qui correspondent à nos critères, il faut pour chacune cliquer sur son lien *Feature link* ce qui vous amène sur l'enregistrement GenBank correspondant. Avant de collecter la séquence dans MEGA, attention à vérifier que l'alignement est sur le Strand=Plus/Plus ou Minus/Minus, sinon il faut aller dans la colonne de droite sous *Customize View* et cocher "Show reverse complement".

Quand vous visualisez les séquences que vous voulez importer, cliquez sur le "+" rouge de la fenêtre de MEGA et la séquence est ajoutée à notre jeu de données.

Question 57 *Construisez ensuite un alignement pour le jeu de données et inférez une phylogénie avec la méthode* Neighbor Joining *après avoir vérifié que le niveau moyen de divergence entre séquences n'excède pas ce qu'elle permet de résoudre correctement (rappel : valeur seuil de 1.0).*