COMO FAZER ALINHAMENTO MULTIPLO E ARVORES FILOGENETICAS

Antes de começar, crie uma pasta onde você colocará todos os arquivos que serão utilizados e gerados nesta aula:

1) Abra o terminal (Ctrl + Alt + T);
2) Crie uma pasta:
>mkdir aula_filogenia
3) Entre na pasta:
>cd aula_filogenia

Programas e Arquivos necessários:

Todos os arquivos e programas necessários para a aula estão disponíveis na página biodados, no link "pinguim", na pasta "filogenia" (http://biodados.icb.ufmg.br/pinguim/filogenia/).

1) Programa de alinhamento:

Existem vários programas que podem ser utilizados para realizar o alinhamento múltiplo de sequências. Dentre eles podemos citar o ClustalW (Larkin et al., 2007), Muscle (Edgar, 2004), Mafft (Katoh et al., 2004), MultiAlin (Corpet, 1988) e o PRANK (Löytynoja e Goldman, 2010) (links no final do texto).

Nesta aula utilizaremos o programa MEGA que é uma plataforma que integra vários programas utilizados para estudos evolutivos. Nele estão implementados o ClustalW e o Muscle. Também utilizaremos o PRANK, que é um alinhador de alta acurácia.

Para quem estiver usando o seu próprio laptop, baixe e instale o MEGA no seu computador. Instaladores para Windows e Mac estão presentes na pasta aula_filogenia ou no site do desenvolvedor (www.megasoftware.net/).

2) Arquivo MULTI-FASTA contendo as sequências a serem alinhadas.

Nesta aula trabalharemos com algumas enzimas que participam na biossíntese ou no metabolismo de alguns aminoácidos. Primeiramente trabalharemos com o arquivo multi-fasta com o nome "*G1.fasta*". Baixe-o na sua pasta usando o wget:

>wget http://biodados.icb.ufmg.br/pinguim/filogenia/G1.fasta

O arquivo "G1.fasta" contém homólogos da enzima "serine—pyruvate aminotransferase", que participa no metabolismo da glicina e serina (veja a rota metabólica no KEGG neste link: http://www.genome.jp/kegg-bin/show_pathway?hsa00260+189).

Alinhamento múltiplo:

A inspeção do alinhamento múltiplo pode revelar diferentes eventos evolutivos, como mutações pontuais, inserções-deleções (INDELS), regiões conservadas, possíveis sítios ativos de uma proteína etc.

Todos os alinhamentos, seja de sequências de nucleotídeo ou de aminoácido, consideram uma matriz de substituição. A matriz de substituição contém a taxa com que um determinado nucleotídeo ou aminoácido altera para outro ao longo do tempo. Os programas de alinhamento utilizam esta matriz para pontuar cada posição do alinhamento e realizar as devidas modificações (considerar match/mismatch, abrir ou estender gap).

O método mais usado no alinhamento múltiplo é o Alinhamento Progressivo, que por ser heurístico é mais rápido, mas mais sujeito a falhas. Consiste em formar diversos alinhamentos par a par que se somam formando o alinhamento global final.

Realizando o alinhamento múltiplo:

1) Abra o programa MEGA (6.06) (ícone com um "M" azul);
2) Na barra de menu, clique em:
   Align → *Edit/Build Alignment*;
   Selecione: *Create a new alignment* (clique em OK) e *Protein* (já que vamos lidar com sequências de proteína)
3) Depois de aberto a janela "Alignment Explorer", importe o arquivo multi-FASTA (G1.fasta). Para isso, copie e cole as sequências em formato FASTA no programa, ou, no menu, vá em *Edit* → *Insert Sequence from File* e selecione o arquivo multi-FASTA;
4) No menu, vá em *Alignment* → *Align by MUSCLE*;
5) Deixe os parâmetros padrão e peça para o programa alinhar;
6) Inspecione o alinhamento que o programa realizou;

Alinhamento múltiplo de sequências utilizando PRANK

1) Entre no site http://www.ebi.ac.uk/goldman-srv/webprank/
2) Copie e cole o seu Multi-FASTA no campo *sequence data*.
3) Na aba *Basic alignment options*, desmarque a opção "*trust insertions (+F)*";
4) Clique em "Start alignment";
5) Espere o alinhamento ficar pronto ou copie o url da página para acessar os resultados posteriormente;
6) Quando o alinhamento estiver pronto, visualize-o em formato FASTA, clicando em "*Show*";
7) Copie todo o alinhamento e cole no Alignment Explorer do MEGA.
8) Para proceder com a análise filogenética, vá em Data → Phylogenetic analysis;

Obtendo a distância entre as sequências:

1) Volte para a janela principal do MEGA;
2) Na barra de menu, clique em:
   Distance → *Compute Pairwise Distance*
3) Deixe nos seguintes parâmetros:
   a. Variance Estimation Method: *None*
   b. Model/Method: *p-distance*
   c. Rates among Sites: *Uniform rates*
   d. Gaps/Missing Data Treatment: *Pairwise deletion*
4) Compute o resultado;
5) Salve a tabela de distância no formato CSV na sua pasta (File → Export/Print distances), salve e visualize no EXCEL ou em qualquer outro programa de planilhas.

Reconstrução da árvore filogenética:

1) Volte para a janela principal do MEGA;
2) Na barra de menu, clique em *Phylogeny*;
3) No MEGA, podemos escolher entre cinco métodos de construção da árvore filogenética:
   a. Neighbor-Joining

b. Minimum Evolution
c. Maximum Likelihood
d. Maximum Parsimony
e. UPGMA

Escolha o *Neighbor-Joining*, que é um método simples de reconstrução de árvore.

4) Configure os seguintes parâmetros:
    a. Test of Phylogeny - *Bootstrap method*
    b. No of Bootstrap replication - *1000*
    c. Model/Method – *p-distance*
    d. Rate Among sites – *Uniform rates*
    e. Gap/missing data treatment – *Pairwise deletion*
5) Inicie e aguarde a análise;
6) Analise a árvore;
7) Salve a árvore no formato NEWICK na sua pasta (File →Export Current Tree (Newick)).

O formato newick pode ser usado como entrada para outros programas que desenham a árvore como Figtree (Rambaut, 2006), TreeView (Page, 1996) e iTOL (Letunic e Bork, 2006) (links no final do texto).

(Opcional) Baixe o arquivo figtree.jar presente na pasta aula_filogenia no servidor pinguim e abra a árvore no formato newick usando este programa. Para isso, siga as intruções:

Abra o programa no terminal pelo comando:

>java –jar figtree.jar

Dentro do programa, abra o arquivo no formato NEWICK (File → open).

Assumimos que as sequências envolvidas no alinhamento possuem uma relação evolutiva e provavelmente possuem um ancestral comum. Caso uma sequência seja muito divergente, o programa poderá não completar a análise.

Explorando as ferramentas do TreeExplorer do MEGA:

O TreeExplorer do MEGA possui várias ferramentas que auxiliam na análise e na estética da sua árvore filogenética.

A) Identificando e colorindo ramos

1) Com o botão esquerdo, clique sobre um ramo da sua árvore para selecioná-la;
2) Com o botão direito, clique sobre o ramo selecionado e vá para “Selected subtree”;
3) Em “Name/Caption” coloque um nome apropriado para o ramo selecionado;
4) Em “Branch Line”, configure a cor, a espessura e o estilo do ramo selecionado.

B) Obtendo o tamanho dos ramos

1) No menu lateral do TreeExplorer, clique em  ;
2) Sobre cada ramo aparecerá um valor que corresponde ao seu comprimento.

C) Diferentes formas de vizualização da árvore

1) No menu superior do TreeExplorer, clique em  ;
2) Nele você encontrará outras opções para vizualizar a sua árvore. Explore todas as formas.

Exercício:
Refaça os procedimentos estudados neste roteiro utilizando desta vez o arquivo *VIL1.fasta*, utilizando o PRANK como alinhador. Após a análise compare as duas árvores geradas (a partir do *G1.fasta* e do *VIL1.fasta*) e analise as diferenças entre elas. Para isso, calcule as seguintes distâncias e complete a tabela abaixo:

	G1.fasta	VIL1.fasta
Animal X Fungo		
Animal X Planta		
Fungo X Planta		

O arquivo "VIL1.fasta" contém homólogos da enzima "acetolactate synthase", que participa na via de biossíntese de valina e de isoleucina (veja a rota metabólica no KEGG neste link: http://www.genome.jp/kegg-bin/show_pathway?ec00290+2.2.1.6).

A rota de biossíntese desses aminoácidos encontra-se completa nos fungos e nas plantas, mas não em animais, o que torna a valina, leucina e a isoleucina aminoácidos essenciais para os últimos. Mas se nós não produzimos estes aminoácidos, qual seria a função desta enzima que faz parte de uma reação intermediária da biossíntese desses aminoácidos nos animais?

A análise filogenética da árvore "VIL1.fasta" indica uma conservação dessas enzimas nos fungos e plantas, e um distanciamento dessa enzima de animais em relação às enzimas de fungos e de plantas. Esse distanciamento indica uma alta taxa de mutação, o que é comum em proteínas que não estejam sofrendo pressão seletiva. Proteínas com esse comportamento pode seguir pelo menos dois possíveis caminhos: (1) desaparecer do genoma, como ocorreu com as outras enzimas que integra a rota de biossíntese desses aminoácidos, ou (2) adquirir uma nova função e fixar na população. Nesta análise, a proteína "acetolactate synthase" dos animais ilustra bem o segundo caso. O próximo desafio seria identificar a nova função que esta enzima adquiriu nesses organismos.

Programas largamente utilizados para realizar Alinhamento Múltiplo:

ClustalW2 - http://www.ebi.ac.uk/Tools/msa/clustalw2/
Muscle - http://www.ebi.ac.uk/Tools/msa/muscle/
MultiAlin - http://multalin.toulouse.inra.fr/multalin/multalin.html
Prank - http://www.ebi.ac.uk/goldman-srv/webprank/
Mafft - http://mafft.cbrc.jp/alignment/server/

Outros programas de vizualização de árvore filogenética:

Figtree - http://tree.bio.ed.ac.uk/software/figtree/
iTOL - http://itol.embl.de/
TreeView - http://taxonomy.zoology.gla.ac.uk/rod/treeview.html