

1ª parte

Análise de sequências mitocondriais do género *Homo*, incluindo DNA de fósseis

Estes exercícios baseiam-se em dois ficheiros com alinhamentos já preparados que acompanham este documento, um da citocromo c oxidase subunidade 1 (COX1/COI), outro do D-Loop, uma região não codificante reguladora da transcrição (<https://doi.org/10.1074/jbc.REV120.011202>) e que é hipervariável.

Nos dois ficheiros encontram-se representados os mesmos genomas mitocondriais, com as seguintes exceções:

- A sequência com a referência AF134583 é dum cromossoma nuclear com origem num genoma mitocondrial ancestral (ver PDF anexo; na altura não se sabia, agora pode ver-se que veio do grupo Denisova), e não contém o D-Loop
- As sequências D-Loop do género *Pan* eram demasiado divergentes e não alinhavam com as do género *Homo*, por isso esse outgroup só serve para as sequências COX1
- Para clarificar a posição da sequência D-Loop MK248454, obtiveram-se 3 sequências adicionais identificadas como pertencendo também ao haplogrupo mitocondrial L0k (<https://doi.org/10.1186/1471-2148-13-56>).

1º exercício

Abrir Homo(COX1).mas (Data>Open a File/Session), determinar o melhor modelo de substituição e realizar a reconstrução filogenética — embora o melhor método seja o da verosimilhança máxima (Maximum Likelihood), podem tentar-se outros, nomeadamente o da evolução mínima (Minimum Evolution) e o da parcimónia máxima (Maximum Parsimony).

Interpretar os parentescos sugeridos nas árvores obtidas. Por exemplo: se se considera *Homo heidelbergensis* uma espécie diferente de *Homo sapiens*, o epíteto *neandarthalsensis* (<https://en.wikipedia.org/wiki/Neanderthal>) deverá ser separado ao nível de espécie ou de subespécie?¹

2º exercício

Repetir o 1º exercício mas considerando apenas a 3ª posição de cada codão. Para isso, no diálogo das preferências de análise, devem desmarcar-se as caixas 1st, 2nd, e Noncoding Sites (secção DATA SUBSET TO USE > Select Codon positions), tanto para a escolha do modelo de substituição como para a filogenia. Compare os resultados (modelos de substituição; filogenias) com os do 1º exercício. Abra de novo o ficheiro Homo(COX1).mas na janela de alinhamento (Align>Open Saved Alignment Session) e verifique como quase todas as substituições no ingroup (género *Homo*) são na 3ª posição.

3º exercício

Repetir o 1º exercício com o ficheiro Homo(D-Loop).mas. Visualize-o na janela de alinhamento para verificar a presença de inserções/deleções (gaps). Dado que não há outgroup, pode ser vantajoso visualizar a árvore “unrooted”: menu View>Tree/Branch style>Radiation.

¹ Obviamente, não é apenas um gene mitocondrial que serve para definir isso, mas com a evidência obtida no exercício pode pensar-se na questão.

2ª parte

Reconstrução da filogenia dos *Mammalia* com base em ortólogos de quitinase 1

Incluem-se para este exercício um ficheiro com o alinhamento já preparado (Chitinase1.meg), e uma filogenia obtida (Chitinase1.png) para servir de referência. Nesta última veem-se chavetas a englobar sequências dum determinado grupo taxonómico, por exemplo *Cetacea*. Isso é conseguido na janela de visualização de dados (botão TA) através dum botão (entre Display e Search) que permite a criação de grupos. O formato meg permite esta gestão da informação, mas o formato mas não.

As chavetas mais externas (*Laurasiotheria*, etc.) são produzidas clicando no respetivo ramo da filogenia e selecionando o menu Subtree>Draw Options>Format Subtree, introduzindo o nome e selecionando OK.

Este alinhamento é muito complexo, com várias posições onde a homologia é fortemente discutível. A grande prevalência de indels de uma ou poucas sequências recomenda que, para poder utilizar mais informação do alinhamento, se proceda do seguinte modo: nas opções de análise, a opção Gaps/Missing Data Treatment (DATA SUBSET TO USE) seja Partial Deletion, com um Site Coverage Cutoff de 95.

O tempo de computação é consideravelmente mais longo que na 1ª parte, por isso é recomendável não colocar demasiadas réplicas do procedimento Bootstrap (ou então deixar o computador a fazer as contas durante a noite).

Adicionais (caso interessem)

Filogenia dos *Tetrapoda*, gene *FOXP4*

<https://www.genecards.org/cgi-bin/carddisp.pl?gene=FOXP4>

Os ficheiros Foxp4_refseq_transl.meg e Foxp4_refseq_transcript_transl.meg contêm sequências de DNA e de proteínas, respetivamente, usando as de 3 anfíbios como outgroup para os *Amniota*. Também existem indels, mas neste caso pode fazer-se a análise com a opção Complete Deletion.

Embora seja informação em base idêntica (isto é, a tradução foi feita a partir do alinhamento de DNA), apenas um deles dá a filogenia correta. Procure interpretar o que pode estar na origem do resultado errado, tendo em conta que são os mesmos genes.

Filogenia dos *Primates*, gene *ESR1*

<https://www.genecards.org/cgi-bin/carddisp.pl?gene=ESR1>

Abrir o ficheiro ESR1_primates_refseq_transcript.fasta primeiro no Alignment Explorer, e determinar os limites (posições nucleotídicas) do *Open Reading Frame* (dica: são 1887 posições). Tomar nota, fechar sem gravar, e em seguida abrir, desta vez no Sequence Data Explorer, onde se podem definir os domínios de análise, com o seguinte procedimento: menu Data > Select & Edit Genes/&Domains..., na linha com fundo azul clicar no botão Edit, abrindo-se a janela Genes and Domain Properties; abaixo de Domain pode escrever-se ORF (ou outro identificador da

sequência codificante), e preencher as posições nas entradas From Site e To Site (em princípio não será necessário alterar as outras opções). Concluir clicando em 2 botões OK. Gravar a sessão (menu Data > Save Session).

A partir deste ORF, pode optar-se por usar toda a informação, ou só as 1ª e 2ª posições de cada codão, ou só a 3ª. Cada uma destas opções requer um modelo de substituição diferente. Observe as diferenças quantitativas (suporte de bootstrap, por exemplo) entre diferentes opções. Note-se uma certa anomalia nos parentescos entre *Hominidae*.

Pode também optar-se por tentar usar a informação do DNA, sem levar em consideração se é na região codificante ou não. Nesse caso, abrir o ficheiro ESR1_primates_refseq_transcript.fasta e trabalhar diretamente com ele. Sugere-se usar Partial Deletion a 60% para aproveitar quase 6000 posições. Aqui, devido à existência de sequências bastante mais curtas que as outras, verifica-se que algumas irão aparecer em posições anómalas se se usar o método Minimum Evolution, mas não no Maximum Likelihood ou no Maximum Parsimony. A anomalia acima referida já não deverá aparecer.