Laboratórios de Bioinformática – 2016/2017

Trabalho prático – enunciado

A *Legionella pneumophila* é uma bactéria gram negativa, do filo Proteobacteria, classe Gammaproteobacteria e ordem Legionellales. Trata-se de uma bactéria patogénica para os seres humanos, que habita essencialmente em reservatórios aquáticos, e que provoca a designada doença do legionário ou legionelose.

O objetivo deste trabalho passa pela utilização das ferramentas bioinformáticas estudadas na unidade curricular de *Laboratórios de Bioinformática* na análise do genoma desta bactéria, usando a sequenciação da estirpe *Legionella pneumophila subsp. pneumophila str. Philadelphia 1* (NCBI taxon: 272624), que contém aproximadamente 3000 genes, num genoma circular com cerca de 3.4 milhões de bases. O registo do NCBI RefSeq com o identificador (Accession) **NC_002942.5** será utilizado para o efeito. A cada grupo de trabalho será atribuída uma zona do genoma, com conjunto aproximado de 200 genes de acordo com a tabela dada em anexo.

O objetivo primordial do trabalho em curso passa pela **anotação funcional** do genoma do organismo de interesse, i.e. a atribuição de funções ao máximo número possível de genes e proteínas do organismo. Para o efeito, devem usar as ferramentas bioinformáticas estudadas na aula, e outras que considerem de interesse, bem como a consulta a bases de dados e literatura (e.g. artigos) relevante. Estas devem ser usadas para validar as anotações originais já disponíveis para as sequências do organismo e procurando complementá-las ou corrigi-las nos casos em que estas sejam inconclusivas, incompletas ou erróneas, fazendo assim a integração da informação disponível.

Para o efeito, os grupos deverão, sempre que possível, desenvolver scripts de análise que possam automatizar as tarefas de forma a tornar possível correr análises para grandes números de genes, sem prejuízo da análise "manual" dos resultados. Em alguns casos, será ainda de considerar a utilização de ferramentas

e pesquisas específicas para genes de maior interesse que não seja possível ou desejável correr para todos os casos.

Os genes e as respetivas proteínas deverão ser caracterizados em termos da sua função (e.g. metabólica – enzimas e transportadores, regulatória, sinalização, etc), sendo também identificadas as bases de dados na pesquisa de informação de interesse, mantendo os respetivos identificadores. Entre as diversas questões biológicas relevantes a abordar no trabalho podem incluir-se a análise do papel dos genes/proteínas atribuídos no processo de infeção e interação com o hospedeiro humano, na forma como os fármacos (e.g. antibióticos) para a doença relacionada atuam e processos de resistência, na aquisição de nutrientes pela bactéria a partir do meio, na invasão de amebas ou na transferência de ADN através do meio aquoso.

Cada grupo deverá criar um sítio web com os resultados do seu trabalho, partilhando os resultados obtidos, na forma de tabelas com as anotações dos genes e respetivas observações, relatórios explicando as análises realizadas e código usado (podendo neste último caso usar serviços específicos para partilha de código como o GitHub). Como forma de ilustrar o uso das scripts desenvolvidas poderão ser usadas as potencialidades dos IPyhton notebooks (http://ipython.org/notebook.html).

Dada a natureza do trabalho, um dos resultados esperados será uma **tabela** (em língua inglesa) onde se resumam as anotações dos genes/proteínas atribuídos a cada grupo. Esta tabela (poderá ter a forma de uma folha de cálculo) deverá conter como informação **mínima** as seguintes colunas:

- Identificação do gene (GeneID NCBI, Accession number NCBI, locus tag, nome do gene, strand)
- Identificação da proteína: Uniprot ID (se disponível) e grau de revisão, Accession number NCBI da proteína, nome da proteína
- Propriedades da proteína: nº de aminoácidos, localização celular
- Lista de termos do GeneOntology (GO) associados ao gene/ proteína
- EC number(s) ou TC number(s) associado(s) (se existirem)

- Descrição: campo de texto livre para a função da proteína
- Comentários (livre, pode ser usado para algum comentário relevante sobre gene ou processo de inferência da função)

Esta tabela poderá ter outros campos que se julguem relevantes e deverá ser complementada por um ficheiro com um "**relatório**" mais documentado sobre cada gene/proteína, que inclua a lista de resultados desse gene e a forma como foi inferida a sua função, bem como a curação manual realizada.
Poderá ainda incluir-se uma classificação da função das proteínas que possa facilmente ser usada para listar proteínas com determinado tipo de função. Como exemplo, são dadas as seguintes classes (não necessariamente exclusivas):

- Metabolismo (enzimas com função metabólica)
- Transportadores (entre compartimentos e para o exterior/interior da célula)
- Regulação (fatores de transcrição e outras proteínas regulatórias)
- Sinalização (proteínas envolvidas nas vias de transdução de sinal)
- Movimento (quimiotaxia, motilidade, fagocitose)
- Processamento do DNA e RNA: reparação, replicação, transcrição, degradação, modificação
- Síntese e processamento de proteínas: tradução, degradação, assemblagem do ribossoma, modificações, encaminhamento
- Outras funções
- Função desconhecida
- Genes codificando RNA (tRNA, rRNA, ncRNA)

Os grupos são **encorajados a colaborar entre si no desenvolvimento de ferramentas de análise**, bem como nos casos onde haja interação entre genes atribuídos a grupos envolvidos em funções biológicas partilhadas. Nos casos de utilização de scripts desenvolvidas por outros grupos, é importante que os créditos sejam claramente identificados.

De forma a orientar os grupos no trabalho sugerindo possíveis abordagens e resultados, este enunciado genérico inicial será complementado por sugestões

de tarefas específicas disponíveis como anexo II a este documento. Note que algumas das sugestões abordam ferramentas que só serão tratadas nas aulas em momento posterior à divulgação deste enunciado, mas ainda em tempo útil para a sua finalização.

Anexo I

Tabela de distribuição do genoma pelos grupos de trabalho

Grupo de trabalho	Genes (locus tag)	Zona do genoma
1	lpg1 a lpg231	1 a 270000
2	lpg232 a lpg462	270001 a 505535
3	lpg463 a lpg693	505536 a 753100
4	lpg694 a lpg924	753101 a 998650
5	lpg925 a lpg1155	998651 a 1275600
6	lpg1156 a lpg1386	1275601 a 1533850
7	lpg1387 a lpg1617	1533851 a 1786150
8	lpg1618 a lpg1848	1786151 a 2072130
9	lpg1849 a lpg2078	2072131 a 2327100
10	lpg2079 a lpg2310	2327101 a 2610800
11	lpg2311 a lpg2541	2610801 a 2873800
12	lpg2542 a lpg2772	2873801 a 3124550
13	lpg2773 a lpg3005	3124551 a 3397754

Anexo II

Sugestões de tarefas

Análise de literatura

Numa primeira fase, deverá procurar alguma literatura genérica que lhe permita conhecer melhor o organismo. Numa fase posterior, poderá procurar artigos específicos para algumas funções biológicas ou genes específicos que possam ajudar a melhorar o seu processo de anotação. A base de dados PubMed poderá

ser de grande ajuda nesta tarefa, podendo as pesquisas ser automatizadas com o Biopython (ver por exemplo secção 9.14.1 do tutorial). Note-se que, dado que se pretende identificar a função de genes individuais, as pesquisas bibliográficas terão que ser em muitos casos específicas, usando-se nas queries nomes de genes/proteínas, processos/funções biológicas, vias metabólicas, etc.

Análise da sequência e das features presentes no NCBI

Deverá desenvolver scripts em BioPython que lhe permitam:

- aceder ao NCBI e guardar o ficheiro correspondente à zona do genoma que lhe corresponde (preferencialmente em ficheiros genbank)
- verificar as anotações correspondentes à zona definida, nomeadamente as do tipo CDS e gene; valide a informação com a tabela presente em: http://www.ncbi.nlm.nih.gov/genome/proteins/416?genome_assembly_id=166758
- verifique e analise toda a informação complementar fornecida pela lista de features e seus qualifiers; note que deve aceder aos registos correspondentes a cada sequência de DNA e proteína para procurar informação adicional; pode ainda usar os campos de referências externas para identificar identificadores de outras bases de dados que permitam solidificar o conhecimento em relação a cada gene

Análise de homologias por BLAST

As ferramentas de procura de homologias serão de especial relevo, requerendo que os resultados obtidos para cada pesquisa sejam analisados procurando inferir pelas sequências homólogas as possíveis funções da sequência original (*query*). Este processo implica analisar a lista de sequências homólogas e identificar padrões consistentes ao nível da função desempenhada por estas. Estes processos deverão ser, sempre que possível, automatizados, mas não se dispensará em muitos casos a análise manual dos resultados.

Poderá desenvolver scripts BioPython para correr a ferramenta BLAST usando como *query* cada uma das sequências (preferencialmente proteínas) atribuídas. Deverá guardar os resultados respetivos e criar scripts para a sua análise semi-automática. Estes poderão ser usados para melhorar a anotação original do

Genbank. Note que para cada sequência irá ter um conjunto alargado de resultados e deverá elaborar e desenvolver estratégias que lhe permitam extrair informação que possa ser automaticamente avaliada. Correr o Blast contra bases de dados mais curadas poderá ser uma hipótese para reduzir o número de resultados e aumentar a sua fiabilidade, mas também poderá dar menos resultados em sequências com pouca homologia.

Ferramentas de análise das propriedades da proteína

Ao longo das aulas da unidade curricular foram estudadas algumas bases de dados e ferramentas que permitem consultar ou inferir algumas das propriedades de uma proteína de interesse.

A base de dados Uniprot permite aceder a toda a informação das proteínas do organismo de interesse. Acedendo pela opção Proteomes pode procurar o proteoma de referência para esta espécie e analisar a informação aí contida (http://www.uniprot.org/proteomes/UP000000609). Os ficheiros da SwissProt podem ser tratados automaticamente pelo BioPython (ver exemplos na secção 10.1 do tutorial).

Note que os registos Uniprot podem ter diferentes graus de revisão por parte dos curadores da base de dados, sendo nos casos em que o registo tenha sido manualmente curado uma fonte importante de informação.

Por outro lado, a base de dados PDB contém informação sobre a estrutura das proteínas. Poderá efetuar pesquisas nesta base de dados no sentido de identificar proteínas do organismo de interesse que estejam presentes nesta base de dados.

Complementarmente, foram estudadas ferramentas que permitem inferir características da proteína com base na sua sequência, como sejam a sua localização celular, a existência de domínios transmembranares ou alterações pós-tradução relevantes. Todas estas ferramentas permitem dar pistas sobre a anotação funcional das proteínas de interesse.

Bases de dados de domínios de proteínas

Nas aulas da unidade curricular foram abordadas bases de dados de domínios de proteínas, das quais se destaca a NCBI CDD (*conserved domain database*) do

NCBI. Esta base de dados, ou outras similares, pode ser usada para confirmar a anotação de proteínas de interesse, sendo de particular utilidade quando subsistem dúvidas sobre a anotação, quer esta provenha da anotação original, quer provenha de resultados de homologia (e.g. BLAST). A CDD permite a pesquisa de proteínas individuais ou pesquisa em batch que podem ser úteis para automatizar processos de procura de conjuntos de proteínas de interesse em simultâneo de forma automática.
Note que, em muitas das proteínas identificadas na anotação há ligações para registos da CDD.

Alinhamento múltiplo e filogenia

As ferramentas estudadas na aula que permitem o alinhamento múltiplo de sequências podem ser úteis no estudo mais aprofundado de alguns dos genes/ proteínas de interesse. Neste caso, pode por exemplo selecionar-se a sequência de interesse do organismo e um conjunto de sequências homólogas (e.g. provenientes de um processo de BLAST) de organismos/ estirpes selecionadas, realizar o seu alinhamento múltiplo e complementarmente determinar a árvore filogenética correspondente. O resultado do alinhamento múltiplo poderá permitir analisar zonas de maior/ menor conservação e conduzir à identificação de domínios conservados de proteínas e permitir dar mais confiança a anotações ou mesmo conduzir a hipóteses ainda não determinadas por outros métodos. Por seu lado, a análise da árvore filogenética poderá levar à identificação de situações de evolução distintas entre genes distintos (e.g. transferência horizontal de genes).

Redes metabólica e regulatória

Um desafio complementar à anotação do genoma será a construção de uma rede metabólica, complementada com interações regulatórias conhecidas. Nesse sentido, o primeiro passo será a compilação das funções metabólicas do organismo e esta tarefa será facilitada pela coleção dos genes anotados com este tipo de função (e seus EC numbers) e sua interligação com a bases de dados KEGG, nomeadamente as reações químicas e vias metabólicas aí definidas. Esta informação poderá ser recolhida numa tabela com colunas: identificadores dos

genes (NCBI, KEGG), lista de EC number(s), KEGG ortholog, KEGG reactions IDs e KEGG pathways (KEGG).

Complementarmente, as proteínas transportadoras podem ser identificadas numa tabela similar com TC numbers no lugar de EC numbers, sendo identificados os compostos (famílias de compostos) transportados.

Finalmente, pode ser compilada uma lista de fatores de transcrição anotados e os genes que são regulados por estas proteínas e sinal da respetiva regulação (ativação ou inibição).

Links para sítios com informação / ferramentas de interesse:

- Base de dados PATRIC - recursos para organismos patogénicos: http://patricbrc.vbi.vt.edu/portal/portal/patric/Home
- KEGG: http://www.genome.jp/kegg/ (código do organismo: *lpn*) – coleção alargada de recursos, com destaque para os voltados para vias metabólicas
- BioCyc e MetaCyc: http://metacyc.org/, http://biocyc.org/, http://biocyc.org/organism-summary?object=LPNE272624
- Ferramenta e base de dados LocTree para previsão sub-celular: https://rostlab.org/services/loctree2, https://rostlab.org/services/loctree3/db/bact/272624_Legionella_pneumophila_subsp.bact.lc3
- Base de dados de transportadores: http://www.membranetransport.org/, http://www.membranetransport.org/all_type_btab.php?oOID=lpne1
- Base de dados de fatores de transcrição previstos: http://www.transcriptionfactor.org/
- Base de dados/ previsão de genes com funções de sinalização (two-component systems) - http://www.p2cs.org/