Bioinformática

Trabalho prático – enunciado inicial

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 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 (**GI: 52840256**). 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 da análise a realizar passa pela caracterização funcional dos genes atribuídos a cada grupo, usando para o efeito as ferramentas bioinformáticas estudadas na aula, bem como a consulta a bases de dados e literatura (e.g. artigos) relevante. 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.

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.

Os genes e as respetivas proteínas deverão ser caracterizados em termos da sua função (e.g. metabólica – enzimas, regulatória, cascatas de sinalização, transportadores, etc), sendo também identificadas as bases de dados na pesquisa de informação de interesse, mantendo os respetivos identificadores.

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).

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 a divulgar durante o tempo de desenvolvimento do projeto pelos docentes durante as aulas da unidade curricular. O primeiro conjunto de tarefas sugeridas está já disponível como anexo II a este documento.

A avaliação do projeto será realizada pela consulta regular aos sítios web desenvolvidos por cada grupo e por uma apresentação final dos resultados a realizar por cada grupo no final do semestre.

Anexo I

Tabela de distribuição do genoma pelos grupos de trabalho

Grupo de trabalho	Genes (locus tag)	Zona do genoma
1	lpg1 a lpg215	1a 248700
2	lpg216 a lpg434	248701 a 472800
3	lpg435 a lpg645	4728001 a 693000
4	lpg646 a lpg859	693001 a 934905
5	lpg860 a lpg1074	934906 a 1171300
6	lpg1075 a lpg1290	1171301 a 1418400
7	lpg1291 a lpg1505	1418401 a 1666810
8	lpg1506 a lpg1720	1666811 a 1919210
9	lpg1721 a lpg1935	1919211 a 2160500
10	lpg1936 a lpg2150	2160501 a 2398540
11	lpg2151 a lpg2365	2398541 a 2670700
12	lpg2366 a lpg2581	2670701 a 2911300
13	lpg2582 a lpg2795	2911301 a 3148910
14	lpg2796 a lpg3005	3148911 a 3397754

Anexo II

Primeiro conjunto de sugestões de tarefas

Análise da sequência e das features presentes no NCBI

Numa primeira fase, 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 pode aceder aos registos correspondentes a cada sequência de DNA e proteína para procurar possível 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 literatura

Numa primeira fase, deverá procurar alguma literatura genérica que lhes 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).

Análise de homologias por Blast

Deverá 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 já 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. Note a existência de registos curados ("reviewed") e não curados, i.e. apenas determinados por ferramentas computacionais. Os ficheiros da SwissProt podem ser tratados automaticamente pelo BioPython (ver exemplos na secção 10.1 do tutorial).

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.