Laboratórios de Bioinformática – 2017/2018

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 integrada de vias metabólicas deste organismo, incluindo a anotação funcional de proteínas e genes de interesse para uma dada via, quer ao nível metabólico, quer ao nível da sua regulação.

Será usada como referência a estirpe *Legionella pneumophila subsp. pneumophila str. Philadelphia 1* (NCBI taxon: 272624), embora possam ser analisadas outras estirpes para comparação. Esta bactéria 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** poderá ser utilizado para consultar o genoma e uma possível anotação do NCBI.

A cada grupo de trabalho será atribuída uma (ou mais) **vias metabólicas** para análise, de acordo com a tabela dada em anexo I. O objetivo primordial passa pela **anotação funcional**, i.e. a atribuição de funções biológicas, a genes e proteínas, relacionados com as vias metabólicas selecionadas e suas interações.

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) relevantes. Estas devem ser usadas para validar as anotações originais já disponíveis para as sequências de genes/ proteínas relacionadas com a(s) via(s) selecionada(s), nomeadamente as presentes no registo do NCBI para o genoma (dado acima) e no BioCyc (ver links das vias para cada grupo).

Devem ainda procurar complementar o conhecimento sobre as vias e os seus genes, fazendo a integração da informação disponível, e.g. em bases de dados e literatura, e dos resultados das ferramentas bioinformáticas. De particular relevância, será tentar complementar as vias preenchendo as lacunas encontradas (i.e. reações putativas sem gene associado), bem como associar as interações regulatórias.

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 estudados por cada grupo 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. Note-se que entre estes genes de interesse podem incluir-se: genes envolvidos diretamente na via codificando enzimas relevantes; genes/proteínas putativamente envolvidos na via, embora sem confirmação experimental; genes/proteínas envolvidos em vias com interações com a via de interesse; genes/proteínas envolvidos na regulação da via de interesse ou em vias de sinalização relevantes nesta regulação.

Deverá ter também uma visão global sobre o papel da via no fenótipo da bactéria e da forma como se relaciona com outras vias em funções biológicas relevantes. 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 estudados por 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.

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 das vias metabólicas pelos grupos de trabalho

Grupo	Via(s) metabólica(s)
1	L-arginine Biosynthesis (1 instances) https://biocyc.org/GCF_001941585/NEW-IMAGE?type=PATHWAY&object=ARGSYN-PWY&detail-level=2

2	L-asparagine Biosynthesis (1 instances) https://biocyc.org/GCF_001941585/NEW-IMAGE?type=PATHWAY&object=PWY490-4 L-aspartate Biosynthesis (1 instances) https://biocyc.org/GCF_001941585/NEW-IMAGE?type=PATHWAY&object=ASPARTATESYN-PWY
3	L-proline Biosynthesis (2 instances) https://biocyc.org/GCF_001941585/new-image?object=PROLINE-SYN
4	peptidoglycan biosynthesis I (meso-diaminopimelate containing) https://biocyc.org/GCF_001941585/NEW-IMAGE?type=PATHWAY&object=PEPTIDOGLYCANSYN-PWY&detail-level=2
5	Quinol and Quinone Biosynthesis (1 via) https://biocyc.org/GCF_001941585/new-image?object=Quinone-Biosynthesis
6	Flavin Biosynthesis (1 instances) https://biocyc.org/GCF_001941585/new-image?object=RIBOSYN2-PWY
7	superpathway of purine nucleotides de novo biosynthesis I https://biocyc.org/GCF_001941585/new-image?object=PWY-841
8	Metabolic Regulators Biosynthesis (1 instances) https://biocyc.org/GCF_001941585/new-image?object=PPGPPMET-PWY
9	L-threonine Biosynthesis (2 instances) https://biocyc.org/GCF_001941585/new-image?object=THREONINE-BIOSYNTHESIS
10	lipid IVA biosynthesis https://biocyc.org/GCF_001941585/new-image?object=NAGLIPASYN-PWY
11	biotin biosynthesis I https://biocyc.org/GCF_001941585/new-image?object=BIOTIN-BIOSYNTHESIS-PWY

Anexo II

https://biocyc.org/GCF_001941585/class-tree?object=Pathways

Análise de literatura

Deverá procurar alguma literatura genérica que lhe permita conhecer melhor o organismo e a via selecionada, bem como artigos específicos para algumas funções biológicas ou genes específicos que possam ajudar a melhorar o seu conhecimento sobre a via e o papel dos genes individuais. 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 da sequência e das features presentes no NCBI

Deverá desenvolver scripts em BioPython que lhe permitam:

- aceder ao NCBI e guardar o ficheiro correspondente ao genoma do organismo filtrando uma lista de genes de interesse para a sua via
- verificar as anotações correspondentes aos genes de interesse, nomeadamente as do tipo CDS e gene; valide a informação com a tabela: 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.

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). Sugere-se também a exploração da análise filogenética para possível comparação de diferentes estirpes da bactéria para genes selecionados.

Regulação das vias metabólicas

Um desafio muito relevante no estudo das vias de interesse será a identificação das interações regulatórias e de sinalização conhecidas que condicionam o funcionamento dessa via. Deve ser compilada uma lista de fatores de transcrição (e outras proteínas regulatórias) anotados com efeitos sobre as vias de interesse, os genes que são regulados por estas proteínas e sinal da respetiva regulação (ativação ou inibição).
Nesta fase, a utilização de ferramentas de procura e de descoberta de motifs podem assumir-se como relevantes.

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/