Laboratórios de Bioinformática
2018/2019

Trabalho prático – enunciado

O objetivo deste trabalho passa pela utilização das ferramentas computacionais estudadas na unidade curricular de *Laboratórios de Bioinformática* na análise integrada do genoma de um organismo patogénico a ser escolhido pelo grupo de trabalho, incluindo a procura de genes homólogos no ser humano e a caracterização de um conjunto de genes selecionados com vista ao seu potencial como potenciais alvos terapêuticos.

O trabalho será dividido nas seguintes fases principais:

- Seleção de um organismo procarionte patogénico, sua caracterização global e da(s) doença(s) que provoca, bem como de tratamentos conhecidos, com foco em fármacos existentes (e.g. antibióticos), através de procuras específicas em literatura e bases de dados relevantes;
- Procura exaustiva por genes homólogos do organismo no ser humano, considerando todos os genes do organismo selecionado, devendo identificar-se a lista dos genes onde estes homólogos não existem; deverão ser usadas ferramentas adequadas para procura de homólogos estudadas na unidade curricular, implementando scripts para a análise automatizada dos seus resultados, definindo critérios adequados;
- Procura por genes considerados essenciais no organismo selecionado por procura em literatura, bases de dados relevantes (e.g. a OGEE) ou outros métodos disponíveis;
- Análise de um conjunto de no mínimo **3** genes considerados essenciais e sem homólogos humanos, e das proteínas que estes codificam, avaliando o seu potencial como alvos terapêuticos; deverá usar procuras em literatura e bases de dados, bem como as diversas ferramentas estudadas fazendo a integração e a interpretação dos diversos resultados obtidos; deverá estender, sempre que relevante, a sua análise a genes relacionados (e.g. na mesma via metabólica, com interações regulatórias, etc) ou a genes homólogos noutros organismos patogénicos. Pelo menos um dos genes selecionado deverá ser um gene ainda não identificado como possível alvo terapêutico, sendo pelo menos um dos outros desejavelmente um alvo já identificado. Bases de dados como o DrugBank ou o CHEMBL podem ser usados para identificar compostos terapêuticos e os seus alvos.

Cada grupo deverá criar um sítio web com os resultados do seu trabalho, partilhando os resultados obtidos, podendo ser incluídos 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). Este sítio web poderá ser atualizado até ao final do dia **15 de janeiro de 2019**.

Os grupos são **encorajados a colaborar entre si no desenvolvimento de ferramentas de análise**. 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 é complementado por sugestões disponíveis como anexo a este documento.

ANEXO:

Sugestões para a execução das tarefas:

Análise de literatura

Deverá procurar alguma literatura genérica que lhe permita conhecer melhor o organismo e os genes selecionados, bem como artigos específicos para algumas funções biológicas que possam ajudar a melhorar o seu conhecimento sobre o papel de genes individuais. A base de dados PubMed poderá ser de grande ajuda nesta tarefa, podendo as pesquisas ser automatizadas com o Biopython.

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;
- verificar as anotações correspondentes a genes de interesse;
- 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, nomeadamente para a procura de genes homólogos humanos, bem como para a caracterização funcional dos genes selecionados. No primeiro caso, deverá configurar adequadamente as suas pesquisas ao nível da base de dados e desenvolver código para automatizar a decisão de existência de homologias significativas. No segundo caso, poderá analisar a lista de sequências homólogas e identificar padrões consistentes ao nível da função desempenhada por estas.

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. 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. As proteínas de interesse podem ser analisadas identificando zonas de possível ligação de compostos que possam regular o seu funcionamento.

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 as proteínas de interesse.

Foram ainda 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). Por outro lado, permite a análise dos domínios

presentes na proteína, de forma a poder caracterizar potenciais pontos de ligação de compostos e outras proteínas que possam inibir o funcionamento da proteína.

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 organismos / estirpes do organismo para genes selecionados.

Regulação

Um desafio muito relevante no estudo dos genes de interesse será a identificação das interações regulatórias e de sinalização conhecidas. Podem ser procurados fatores de transcrição (e outras proteínas regulatórias) anotados com efeitos sobre os genes de interesse, os genes que são regulados por estas proteínas e sinal da respetiva regulação (ativação ou inibição). Por outro lado, os genes de interesse podem ter efeitos regulatórios, individualmente ou por interações com outros genes, condicionando a expressão de outros genes.

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/– coleção alargada de recursos, com destaque para os voltados para vias metabólicas
- BioCyc e MetaCyc: http://metacyc.org/, http://biocyc.org/
- Bases de dados de fármacos e interações: https://www.drugbank.ca/, https://www.ebi.ac.uk/chembldb/, https://pubchem.ncbi.nlm.nih.gov/

- Bases de dados de essencialidade: http://ogee.medgenius.info/browse
- Ferramenta e base de dados LocTree para previsão sub-celular: https://rostlab.org/services/loctree2
- Base de dados de transportadores: http://www.membranetransport.org/
- Base de dados de fatores de transcrição previstos: http://www.transcriptionfactor.org/