Laboratórios de Bioinformática
2020/2021

Trabalho prático – enunciado

O objetivo deste trabalho passa pela **utilização das ferramentas computacionais** estudadas na unidade curricular de *Laboratórios de Bioinformática* **na análise de um conjunto de genes de interesse relacionados com a COVID-19**.

Esta doença é causada por um coronavírus, da família Betacoronavírus, que terá ultrapassado a barreira das espécies e infetado humanos. Este vírus, chamado SARS-CoV-2, em poucos meses espalhou-se por todo o mundo causando até 25 de Novembro de 2020, perto de 60 milhões de infeções confirmadas e de 1,4 milhões de mortes (https://covid19.who.int/).
Muitos esforços estão a ser dirigidos para a investigação desta doença a nível mundial. Cerca de 23000 artigos estão já indexados na base de dados de artigos em investigação médica Pubmed desde o início do ano. Isto indica uma média superior a 2000 artigos científicos publicados por mês acerca desta doença desde o começo do ano, apenas nesta plataforma.
A COVID-19 apresenta uma vasta variabilidade de sintomatologia e severidade da doença. Esta variabilidade encontra-se relacionada com vários fatores que incluem idade, co-morbilidades, etnia/ancestralidade e perfil genético. A procura de genes associados a diferentes suscetibilidades ou proteção à doença é uma fonte importante de investigação no combate à pandemia e vários genes foram já apontados.

Neste trabalho, pretende-se analisar um conjunto de genes, quer a nível do vírus, quer a nível do hospedeiro humano, focando no estudo das sequências associadas e sua estrutura. Os genes selecionados são genes/ proteínas do vírus, bem como genes humanos que estejam relacionados por interações com estes e/ ou envolvidos em processos de infeção. O principal objetivo será a utilização das ferramentas bioinformáticas e interpretação dos seus resultados para perceber, dentro do possível, a função dos genes em causa, suas possíveis interações, suas relações com a doença e com outros fatores genéticos ou ambientais. Note que relacionado com cada gene selecionado poderá haver diversas sequências de interesse (e.g. DNA, proteína, RNA), bem como informação relevante em sequências relacionadas (e.g. homólogas) e informação complementar de interesse em bases de dados e literatura.

Cada grupo realizará a análise de um total de **3 (ou 4)** genes atribuídos, e das proteínas que estes codificam. Cada grupo deverá escolher um dos genes do vírus, bem como um conjunto de 2 ou 3 genes humanos que com este estejam relacionados. Na tabela anexa ao enunciado são listados os principais genes do vírus, bem como sugeridos alguns genes humanos que podem ser interessantes estudar. Os grupos terão liberdade para, com base numa análise bibliográfica prévia, sugerir outros genes alternativos.

No desenvolvimento do trabalho, deverão usar procuras em literatura e bases de dados para caracterizar os genes selecionados e suas funções, com particular ênfase na interação entre genes do vírus e do hospedeiro e sua relação com processos de infeção/ patogenicidade, bem como utilizar as diversas ferramentas bioinformáticas estudadas na unidade curricular (ou outras que considere relevante), desenvolvendo scripts, fazendo a integração e a interpretação dos diversos resultados obtidos. Poderão ainda tentar explorar o potencial de algum dos genes como alvo terapêutico e estender, sempre que relevante, a sua análise a genes relacionados (e.g. na mesma via, com interações regulatórias, etc) ou a genes homólogos noutros organismos.

O trabalho decorrerá de acordo com as seguintes fases:

- **escolha dos genes**: deverão indicar os seus genes no site de e-learning no link devido para se poder validar a escolha; esta fase terá que estar concluída até ao dia **1 de dezembro de 2020**. Os grupos terão 3 ou 4 elementos, tendo que escolher um número de genes igual ao número de elementos do grupo. Deverá submeter um ficheiro PDF com a constituição do grupo e a indicação dos genes escolhidos, podendo colocar várias alternativas de conjuntos de genes, de forma ordenada, para que se possa evitar sobreposições nas escolhas dos grupos.

- ***sítio web***: 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 *Jupyter Notebooks.* Os sites serão avaliados, por consulta dos docentes, em duas datas: uma nos dias 21/ 22 de dezembro (atualização até **20 dezembro 2020**) e a segunda nos dias 1 e 2 fevereiro (atualização até **31 de janeiro 2021**).

- **apresentação intermédia**: apresentação dos resultados dos trabalhos (cerca de 15 minutos por grupo) a realizar na 1ª semana de janeiro (data a definir). Deve focar a metodologia seguida, os resultados obtidos até ao momento e possíveis linhas para trabalho ainda a realizar. Permitirá aos grupos obter feedback para a melhoria do trabalho até ao prazo final.

A avaliação do trabalho será realizada com base na consulta dos sites web nos dois pontos temporais referidos e avaliação dos seus conteúdos (2/3) e pela apresentação intermédia realizada (1/3). Os elementos dos grupos poderão ser avaliados de forma distinta, se para tal os docentes considerarem haver justificação.

Os sites de cada grupo deverão obrigatoriamente incluir uma secção de "Créditos", onde expliquem com clareza os contributos de cada elemento do grupo nas várias tarefas. Os grupos são **encorajados a colaborar entre si no desenvolvimento de ferramentas de análise**. Nestes casos, quando haja a utilização de scripts desenvolvidas por outros grupos, é importante que os créditos sejam claramente identificados nesta mesma secção.

De forma a orientar os grupos no trabalho, sugerindo possíveis abordagens e resultados, este enunciado genérico é complementado pelas linhas orientadoras que se seguem.

Orientações para a execução das tarefas:

Análise de literatura

Deverá procurar alguma literatura genérica que lhe permita conhecer melhor os genes selecionados, bem como artigos específicos para algumas funções biológicas que possam ajudar a melhorar o seu conhecimento sobre o seu papel, quer em casos normais quer em fenótipos de cancro. 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 os ficheiros correspondentes aos genes escolhidos, podendo explorar possíveis variantes;

- verificar as anotações correspondentes aos genes de interesse;
- verificar e analisar a informação complementar fornecida pela lista de *features* e seus *qualifiers*; pode 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, 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. Deverá implementar scripts Python/ BioPython para automatizar estas tarefas.

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 de um conjunto alargado de proteínas. 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 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 selecionados, 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. Sugere-se também a exploração da análise filogenética para possível comparação de diferentes organismos do organismo para genes selecionados. Este processo deve ser realizado para os genes de interesse, idealmente automatizando com BioPython.

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.

Comparação de variantes mutagénicas do gene e o seu impacto biológico

A existência de variantes mutagénicas possibilita a existência de cadeias de transmissão com um perfil genético distinto. Sabe-se que, por exemplo, uma variante pode causar a substituição

de um aminoácido numa proteína com impacto a nível funcional ou estrutural da proteína. Porém, até ao momento, ainda não se conhecem os impactos causados pelas variantes mutagénicas que ocorrem ao longo da evolução do vírus SARS-CoV-2. A base de dados GISAID (https://www.gisaid.org/) é uma iniciativa que tenta recolher e fornecer todos os possíveis variantes identificados em amostras de SARS-CoV-2, tendo disponíveis as sequências de mais de 200000 genomas do vírus. Com base nas variantes obtidas das amostras de SARS-CoV-2 utilizadas na TBL de NGS de Laboratórios de Bioinformática e/ou com variantes identificadas na base de dados GISAID, é possível inferir sobre uma possível alteração biológica do gene em estudo (quer seja funcional ou estrutural) a partir de uma determinada variante. Sugere-se a comparação de variantes em diferentes amostras, por exemplo as que ocorram em países distintos.

Anexo: tabela de genes do vírus SARS-CoV-2 e genes humanos relacionados

Gene - virus	Sequência Uniprot	Proteínas virus	Genes humanos interessantes	
Replicase (ORF1a)	https://www.uniprot.org/uniprot/P0DTC1	NSP2	PHB	https://www.uniprot.org/uniprot/P35232
			PHB2	https://www.uniprot.org/uniprot/Q99623
Spike - S (ORF2)	https://www.uniprot.org/uniprot/P0DTC2	S1	ACE2	https://www.uniprot.org/uniprot/Q9BYF1
			TMPRSS2	https://www.uniprot.org/uniprot/O15393
		S2	PRSS1	https://www.uniprot.org/uniprot/P07477
		Glycoprotein	BST2	https://www.uniprot.org/uniprot/Q10589
ORF3	https://www.uniprot.org/uniprot/P0DTC3		FGA	https://www.uniprot.org/uniprot/P02671
			FGB	https://www.uniprot.org/uniprot/P02675
			FGG	https://www.uniprot.org/uniprot/P02679
			IFNAR1	https://www.uniprot.org/uniprot/P17181
VEMP	https://www.uniprot.org/uniprot/P0DTC4	E	MPP5	https://www.uniprot.org/uniprot/Q8N3R9
			CRB3	https://www.uniprot.org/uniprot/Q9BUF7
VME1	https://www.uniprot.org/uniprot/P0DTC5	M	RTN4	https://www.uniprot.org/uniprot/Q9NQC3
			Q4KMQ2	https://www.uniprot.org/uniprot/Q4KMQ2
ORF6	https://www.uniprot.org/uniprot/P0DTC6		STAT1	https://www.uniprot.org/uniprot/P42224
			KPNA2	https://www.uniprot.org/uniprot/P52292
ORF7a	https://www.uniprot.org/uniprot/P0DTC7		BST2	https://www.uniprot.org/uniprot/Q10589
			ITGAL	https://www.uniprot.org/uniprot/P20701
ORF7b	https://www.uniprot.org/uniprot/P0DTD8		UN93B	https://www.uniprot.org/uniprot/Q9H1C4
			MAVS	https://www.uniprot.org/uniprot/Q7Z434
Nucleoprotein	https://www.uniprot.org/uniprot/P0DTC9	N	SMAD3	https://www.uniprot.org/uniprot/P84022
			HNRNPA1	https://www.uniprot.org/uniprot/P09651
ORF9b	https://www.uniprot.org/uniprot/P0DTD2		MAVS	https://www.uniprot.org/uniprot/Q7Z434
			XPO1	https://www.uniprot.org/uniprot/O14980
			TRAF3	https://www.uniprot.org/uniprot/Q13114