# *Laboratórios de Bioinformática*
# *2022/2023*

## 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 potencialmente relacionados com o diabetes tipo 2.**

O diabetes tipo 2 é uma das mais relevantes doenças que afetam atualmente a humanidade, com impacto social e económico enorme, nomeadamente ao nível da diminuição da esperança e da qualidade de vida das populações e dos efeitos nos sistemas de saúde em todo o mundo. A doença está relacionada com diversos fatores ambientais, mas possui também uma base genética, em larga medida desconhecida.

Neste trabalho, pretende-se analisar um conjunto de genes potencialmente relacionados com esta doença, focando no estudo das sequências associadas e sua estrutura. Os genes selecionados são genes identificados como potencialmente associados à doença em estudos estatísticos de associação de genomas a fenótipos de doenças (GWAS). 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 e suas relações com a doença e com os possíveis fatores ambientais relacionados. Note que relacionado com cada gene selecionado poderá haver diversas sequências de interesse (e.g. DNA, proteína, RNA e suas variantes), 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** genes escolhidos de entre aqueles identificados por estudos de GWAS como potencialmente relevantes, e das proteínas que estes codificam. Os genes deverão ser escolhidos de entre aqueles identificados num estudo recente, nomeadamente os presentes na Tabela 2 do Material Suplementar da seguinte publicação: https://www.nature.com/articles/s41588-018-0241-6. Sempre que possível poderá escolher analisar 3 genes que possuam alguma relação entre si.

No desenvolvimento do trabalho, deverão usar procuras em literatura e bases de dados para caracterizar os genes selecionados e suas funções, 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 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 genes escolhidos no site de e-learning no link devido para se poder validar a escolha; esta fase terá que estar concluída até ao dia **18 de novembro de 2022**. Os grupos terão 3 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 mais do que uma alternativa, de forma ordenada, para que se possam evitar sobreposições nas escolhas dos grupos.

- ***portfolio digital***: cada grupo deverá criar um portfolio digital com os resultados do seu trabalho, usando a aplicação Padlet (www.padlet.com), partilhando os principais resultados obtidos. Será criado um padlet global dos trabalhos, onde cada grupo colocará um resumo do seu trabalho indicando os genes analisados e os *highlights* dos resultados, um link para o seu próprio padlet e um link para o seu repositório (ver abaixo).

- ***repositório***: Deverão ser incluídos relatórios detalhados explicando as análises realizadas e o código usado através do uso do serviço GitHub para partilha de código, dados e relatórios de análise. Como forma de ilustrar o uso das scripts desenvolvidas poderão ser usadas as potencialidades dos *Jupyter Notebooks.*

Os portfolios/ repositórios serão avaliados, por consulta dos docentes, em duas datas: uma no dia **19 de dezembro 2022** e a segunda no dia **3 de fevereiro de 2023** (atualização do site poderá ser realizada até às 9 h. do dia referido).

- **apresentação**: apresentação dos resultados dos trabalhos (cerca de 15 minutos por grupo) a realizar a **25 de janeiro de 2023**. 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 portfolios/ repositórios nos dois pontos temporais referidos e avaliação dos seus conteúdos (2/3) e pela apresentação realizada (1/3). Os elementos dos grupos poderão ser avaliados de forma distinta, se para tal os docentes considerarem haver justificação. Cada elemento do grupo será chamado a dar feedback sobre o trabalho dos colegas de grupo. Cada grupo será chamado a dar feedback sobre a apresentação dos restantes grupos.

Os portfolios 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 ou Diamond**

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. Poderá 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 do gene e o seu impacto biológico**

A existência de mutações poderá estar intimamente relacionada com os fenótipos associados a vários genes. Assim, recomenda-se que possam estudar, com base nas sequências e dados disponíveis em bases de dados, diferentes variantes dos genes selecionados, se tal informação existir.