Universidade do Minho

Mestrado Integrado em Engenharia Informática

Bioinformática - 4º ano

Laboratórios de Bioinformática

Treponema pallidum subsp. pallidum str. Nichols

Cláudia Ribeiro - A64460
Duarte Moças - A64319
Jorge Ferreira - A64343

29 de Janeiro de 2016

Conteúdo

1 Introdução

Treponema pallidum subsp. pallidum Strain Nichols é uma bactéria causadora da doença sexualmente transmissível sífilis, a qual todos os anos provoca novos portadores da mesma. Com o aparecimento e distribuição em larga escala de penincilina, na década de 1940, a transmissão da doença foi bastante mitigada, até à década de 1980-1990. Com o virar do milénio, verificou-se um crescimento das taxas de infeções, em particular nos Estados Unidos da América, Europa, Reino Unido, Canadá e Austrália, sendo que a transmissão da doença deve-se, maioritariamente, à prática de sexo inseguro.

Neste trabalho pretende-se expor as funcionalidades e características dos genes localizados entre a posição 1015201 e 1138011 do genoma do organismo. Para tal, faremos uso das diferentes bases de dados biológicas estudadas, pesquisando por funções, características, localização sub-celular, entre outros atributos, aliando também esta pesquisa ao desenvolvimento de *scripts* em Biopython.

2 Desenvolvimento

De forma a uniformizar a pesquisa de informação, iniciámos o desenvolvimento do trabalho definindo, primeiramente, os campos (colunas da tabela) referentes a cada gene/-proteína obtido.

Para tal, verificámos quais as informações que cada base de dados biológica oferece, de modo a definir uma estrutura de pesquisa. As bases de dados biológicas usadas foram o NCBI[1], UniProt[2] e GeneOntology[3]. A lista seguinte apresenta os campos associados a cada uma das bases de dados:

2.0.1 NCBI

- Accession: identificador único de uma sequência;
- Start (localização inicial);
- Stop (localização final);
- Strand;
- GeneID: identificação único do gene;
- EC Number (Enzyme Comission Number): classificação numérica de enzimas, baseada nas reacções químicas que estas cataliza;
- Locus: localização/posição do gene na sequência;
- Locus Tag;
- Protein Product;
- Nº de Aminoácidos;

2.0.2 UniProt

- UniProtID: identificador único para cada entrada na base de dados UniProt;
- Revision;
- Subcellular Location;
- Classification of Protein Function (Função da Proteína);

[1] http://www.ncbi.nlm.nih.gov/
[2] http://www.uniprot.org/
[3] http://geneontology.org/

2.0.3 GeneOntology

- Protein Function Information: informação sobre a função da proteína;
- GeneOntology Terms;
- GeneOntology Identifiers;

Adicionalmente, existem dois campos, **Description** e **Commentary**, sendo ambas para informações extra ou de síntese.

2.1 Recolha Manual e Uso de *BioPython*

Inicialmente, devido às dificuldades encontradas, recolhemos a informação do genoma de forma manual, pesquisando e acolhendo dados das diferentes bases de dados mencionadas anteriormente. Contudo, após a recolha manual de dados, decidimos desenvolver uma ferramenta que fosse capaz de realizar parte desde processo, de forma automática.

Com isto, desenvolvemos um *script* em *Python*, através do uso da biblioteca *BioPython*[4], capaz de recolher informações das bases de dados *NCBI* e *UniProt*. Através do *Acession Number*, o *script*, primeiramente, transfere a informação disponível no *NCBI*, em formato *genbank*, sendo que apartir deste podemos retirar as informações relevantes. Através do *GI_Number* de cada gene, podemos então aceder ao *UniProt* à pesquisa de organismos que contenham aquele gene. Com isto, extraimos o *UniProtID*, que depois usamos para transferir a página em formato *xml*. A partir deste ficheiro, podemos, então, retirar a restante informação.

Existem, no entanto, informações que não conseguimos obter de forma automática, como a *Função de Proteína* e sua informação adicional.

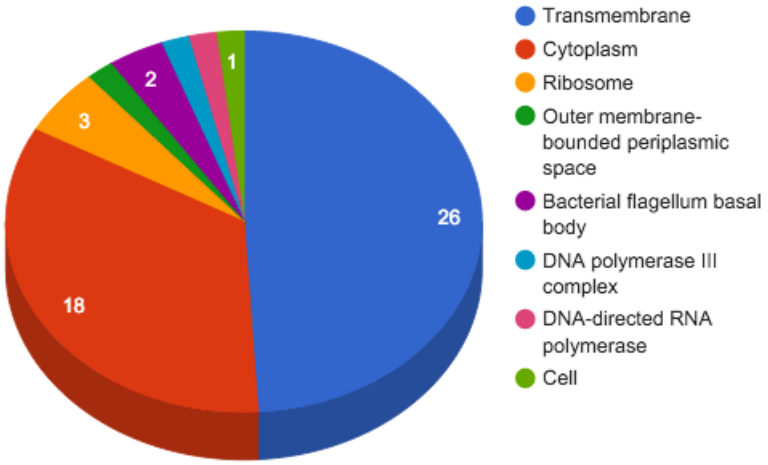
3 Resultados

Os resultados obtidos encontram-se, em formato Excel, neste link, assim como num website, contendo essa mesma tabela.

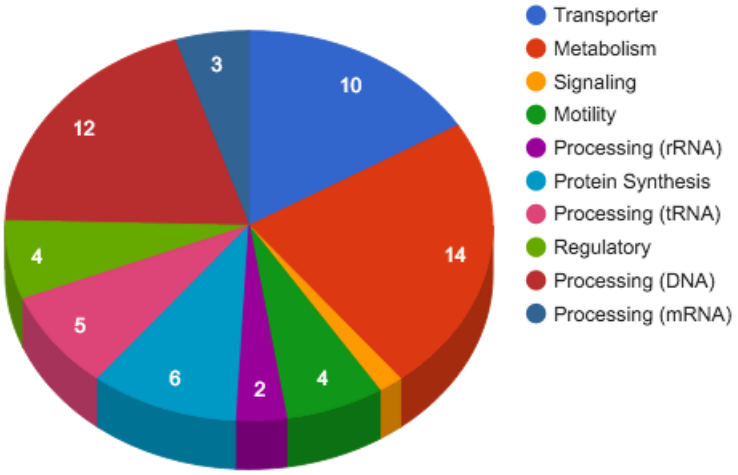
Através dos resultados fizemos estatísticas, por exemplo, da classificação da função da proteína, como é possível ver de seguida:

[4]http://biopython.org/

Subcellular Location



Classification of Protein Function



4 Conclusão

Com a realização deste projecto foi possível aprofundar o conhecimento das ferramentas bioinformáticas utilizadas nas aulas, uma vez que foram necessárias para obter de informação sobre o genoma da bactéria *Treponema pallidum pallidum*, mas especificamente da estirpe *Treponema pallidum subsp. pallidum str. Nichols*.

As principais dificuldades que sentimos foram em iniciar o projecto um pouco devido à falta de conhecimento na área da biologia. Mais tarde, tivemos alguns problemas em construir uma forma automática para a obtenção dos dados, no entanto e tal como já foi demonstrado anteriormente, conseguimos ultrapassar esse obstáculo, tendo sido desenvolvido um *script* em Biopython para esse efeito.

Embora tenhamos conseguido fazer a anotação funcional do genoma, não fizemos o alinhamento múltiplo e filogenia, assim como as redes metabólica e regulatória. Para o alinhamento múltiplo começámos a desenvolver um *script* em Biopython, mas este revelou-se ineficaz e portanto não o incluímos nas soluções apresentadas. No que diz respeito às redes metabólica e regulatória, através de uma navegação sobre a nossa tabela de resultados podemos concluir que a zona do genoma que nos foi atribuída tem enzimas de várias "famílias", isto é, enzimas com EC Number 1, 2, 3 e 6, ou seja, *oxidoreducatases*, *transferases*, *hydrolases* e *ligases*, respectivamente.