Faculdade de Ciências da Universidade de Lisboa

Bioinformática

Bioinformática 2023/2024

Projecto Final

Grupo 10

Nomes	Números
André Borgas	60436
Cristiana Gonçalves	56036
João Vedor	56311
Miguel José	56118

Lisboa, 26 de Maio de 2024

Nota para a Professora:

Nós temos um ficheiro project_notebook com todos os passos efetuados, que deve ser utilizado apenas para ter uma ideia geral do projeto. Os 3 ficheiros python que devem ser corridos, com o código de cada goal em cada um deles, são respectivamente: goal1.py, goal2.py e goal3.py presentes na pasta project_goals.

1. **Select two species and retrieve the respectives genomes.**
    a. **Description of the species:**

Os Homo sapiens, ou seres humanos, possuem um genoma composto por aproximadamente 3 biliões de pares de bases de DNA, distribuídos em 23 pares de cromossomas. Apesar de compartilharem 99,9% do DNA entre si, as pequenas variações genéticas resultam em diferenças individuais significativas.

O genoma humano contém genes que regulam funções essenciais como desenvolvimento cerebral, sistema imunológico e metabolismo. A expressão génica é controlada por uma complexa rede de interações e processos epigenéticos. Polimorfismos genéticos, como SNPs (single nucleotide polymorphisms), contribuem para a diversidade e influenciam a susceptibilidade a doenças e respostas a medicamentos.

Elementos transponíveis, que representam cerca de 45% do genoma, desempenham um papel na evolução e regulação genética. Estudos genómicos também revelam a história evolutiva humana, incluindo migrações e interações com Neandertais e Denisovanos. Resumindo, o genoma humano define tanto as características biológicas quanto a diversidade e adaptação dos Homo sapiens ao longo do tempo.

https://www.ncbi.nlm.nih.gov/nuccore/568336023 - Ser humano

O coronavírus é uma família de vírus que pode causar doenças em animais e humanos. A variante mais recentemente identificada é o SARS-CoV-2, que causa a doença conhecida como COVID-19. Os coronavírus são vírus encapsulados de RNA de cadeia simples, com um genoma de RNA de sentido positivo, ou seja, corresponde a uma cadeia de RNA que tem a mesma sequência de bases que o RNA mensageiro (mRNA) que seria traduzido em proteína. Por outras palavras, o RNA de sentido positivo contém as instruções diretas para

a síntese de proteínas. O seu nome deve-se à sua aparência sob microscópio eletrónico, que apresenta uma coroa com projeções na sua superfície.
Desde o seu surgimento em dezembro de 2019 na cidade de Wuhan, na China, o coronavírus espalhou-se rapidamente pelo mundo, desencadeando uma pandemia global.

https://www.ncbi.nlm.nih.gov/nuccore/MN908947 - Corona virus

b. Description of the genome files: number of records, IDs, names, and description of each record.

O código analisa um ficheiro FASTA de genomas, extraindo e imprimindo informações detalhadas para cada registo, incluindo identificador único (ID), nome e descrição.
Primeiramente, o código conta o número total de registos, armazenando esse valor na variável count_records, que é impressa no final para fornecer uma visão geral do volume de dados. Para cada registo, o código extrai o ID (atributo id do objeto SeqRecord), o nome (name) e a descrição (description), imprimindo esses detalhes para cada sequência.
Embora o comprimento das sequências também seja calculado inicialmente, esses valores não são apresentados de forma a cingir os resultados ao que foi pedido. Em resumo, o código processa o ficheiro FASTA, fornecendo uma análise detalhada de cada registo e contando o número total de registos.
A análise dos dados mostra uma diferença significativa entre o número de registos dos dois genomas analisados: o genoma humano possui uma grande variedade de regiões genómicas, enquanto o genoma do coronavírus contém apenas um registo.

2. **Sequence alignment of the genomes, and calculation of the similarity between the species**
   a. Select one record from each species

Copiámos o primeiro ‘record’ do genoma humano que obtivemos no objetivo número 1 e colocámo-lo num novo FASTA file. Quanto ao genoma do coronavírus, tendo em conta que apresenta um só record, utilizámos esse ‘record’.

b. Alignment of the sequences of the genes using the alignment tools, identification of similarities between the sequences

O alinhamento de sequências é uma ferramenta poderosa para estudar relações evolutivas entre diferentes espécies. Neste caso, o 'record' do SARS-CoV-2 e um 'record' do cromossoma humano 15 forneceu insights sobre a relação entre o vírus e o hospedeiro humano. Após a seleção dos records e o uso de BLAST entre as sequências, diversos indicadores foram analisados para avaliar a similaridade entre as sequências.

Esses indicadores incluem o comprimento de cada sequência, o E-value, que estima a probabilidade de encontrar por acaso um alinhamento com uma pontuação tão alta, e o Score, que indica a qualidade geral do alinhamento. Além disso, o número de identidades representa os nucleotídeos idênticos entre as sequências, enquanto a pontuação de similaridade mostra o quão semelhantes são as sequências. Temos também os gaps, que representam inserções ou deleções introduzidas para melhorar o alinhamento. Por fim, a correspondência visual oferece uma representação gráfica do alinhamento, mostrando como os nucleotídeos se alinham entre as sequências.

E-values e Score: E-values oscilam entre 0.139 e 6.3 e o Score vai de 15 a 18

Através da identidade e similaridade concluímos que existem regiões parecidas. identidade - 15/15 e similaridade de 86.6% a 100%.

Os resultados do alinhamento mostram que certos segmentos do genoma do SARS-CoV-2 têm uma elevada semelhança com regiões do cromossoma humano 15. A elevada identidade e semelhança em segmentos curtos sugere que pode haver sequências conservadas ou correspondências coincidentes, mas os E-values geralmente mais elevados para a maioria dos alinhamentos, implicam que estes resultados não sejam biologicamente significativos. Estes resultados devem ser interpretados com cautela e no contexto de uma análise biológica mais aprofundada para determinar se estas semelhanças têm quaisquer implicações funcionais ou evolutivas.

3. Retrieval of research papers related with each species

a. Biopython tutorial: Chapter 12 - Accessing NCBI's Entrez databases

Neste passo, seguindo as diretrizes no capítulo 12 do Tutorial de Biopython, acedemos à PubMed, mais especificamente à sua biblioteca de informação. O seu acesso é possível através da função search_pubmed_from_term, que recebe um termo para pesquisar como entrada e retorna uma lista com 10 resultados de pesquisa na PubMed. A função usa o módulo Entrez para realizar a pesquisa. O módulo Entrez é uma biblioteca Python que fornece acesso à interface de programação de aplicativos (API) do National Center for Biotechnology Information (NCBI). Por último, a função usa o método read() para ler os resultados da pesquisa. Os resultados são armazenados em uma lista e por fim, a função fecha o fluxo de entrada e retorna a lista de resultados da pesquisa.

b. For each specie, retrieve the 10 most recent research papers from Pubmed

Artigos relacionados com o Homo Sapiens:

Os resultados dos 10 artigos encontrados vão de acordo com o expectável, estando todos relacionados com o genoma humano. No entanto, foram encontrados artigos com data posterior à atual, indicando que o lançamento do artigo está planeado e que o abstrato já foi criado e disponibilizado para a comunidade científica.

Artigos relacionados com o CoronaVírus:

Um dos artigos que nos foi retornado tem o título: "Addressing cortex dysregulation in youth through brain health check coaching and prophylactic brain development", de Blum K. Depois de ver que o título não mostra uma relação direta entre o artigo e coronavírus, lemos o abstract e descobrimos uma possível razão pela qual obtemos este artigo: "In recent years, the well-being of youth has been compromised by not only the coronavirus disease 2019 pandemic but also the alarming global opioid crisis". Embora demonstre uma relação entre o vírus e o seu conteúdo, este não é um artigo relevante para nós. Fomos procurar as keywords e não havia qualquer menção do coronavírus. Como tal, uma sugestão para melhorar a nossa pesquisa seria procurar as keywords de cada artigo para estabelecer uma relação mais evidente em vez de procurar apenas as palavras no texto. Encontrámos mais exemplos como este descrito acima.

Nas datas de alguns artigos não estava presente dia e mês, apresentando por vezes apenas o ano de publicação.