

Mestrado em Bioinformática

**Base de Dados para o armazenamento e informação contida em ficheiro de formato Genbank**

Autores:

* José Lemos
* Paulo Seixal
* Rúben Fernandes

Introdução aos Algoritmos, à Programação e às Bases de Dados

Janeiro de 2023

**Índice**

[Introdução 3](#_Toc124083346)

[Motivação e objetivos 3](#_Toc124083347)

[Modelo conceptual 3](#_Toc124083348)

[Modelo lógico e modelo físico 3](#_Toc124083349)

[*Functions*, *procedures* e *triggers* 3](#_Toc124083350)

[Povoamento das tabelas 3](#_Toc124083351)

[Discussão e análise 3](#_Toc124083352)

[Bibliografia 3](#_Toc124083353)

# Introdução

O que é um ficheiro GenBank:

Os ficheiros GenBank permite o armazenamento de mais informação para além da sequência.

- O que contem/objetivo:

-file format:

O formato dos ficheiros GenBank são compostos, principalmente por:

* **LOCUS** – Armazena informação básica do registo.
  + Nome do Locus (ex: SCU49845)
  + N.º pares de bases
  + Tipo de sequência
  + Direção da sequência
  + Accession?
  + Divisão do GenBank
  + Data de submissão
* **DEFINITION** – Breve descrição da sequência onde menciona o organismo, nomes dos genes/proteínas.
* **ACCESSION** – Indentificador único para o registo da sequência
* **VERSION** – Representado por um complexo de duas partes onde o primário é composto pelo accession e um número de versão representativo de uma alteração da sequência (ex.: U49845.1), a segunda chave é designada pela palavra-chave GI seguida de um número que é alterado quando é efetuada uma alteração na sequência.
* **KEYWORDS** – Palavra ou frase que descreve a sequência. Registo sem keywords são representados por um ponto final. Estas não são incluídas em novos registos a não ser que: (1) – Não sejam redundantes com qualquer feature, qualifier ou outra informação; (2) – O responsável pela submissão faça um pedido; (3) – O registo tenha um tipo especial de sequência.
* **SOURCE** - Indica o nome mais comum usado para o organismo de onde a sequencia foi retirada.
  + ORGANISM – Indica o nome científico e a sua taxonomia.
* **REFERENCE** – Campo onde se referencia citações e publicações de origem da informação presente no ficheiro. Dentre deste campo também se pode consultar:
  + AUTHORS
  + TITLE
  + JOURNAL
  + PUBMED
* **FEATURES** – Contém informação de interesse para os investigadores, localização da porção de uma sequência e informação relativa a essa porção.
  + source
  + CDS
  + gene
* **ORIGIN** – Contém a sequência em estudo.

## Motivação e objetivos

Foi proposto como elemento de avaliação a elaboração de um trabalho que inclui a resolução de um problema existente nos ficheiros GenBank. De forma a solucionar este problema, foi proposta a criação de uma base de dados para transferir, organizar e armazenar informações referentes a ficheiros GenBank.

Os ficheiros em formato GenBank armazenam muita informação associada a várias sequencias biológicas de vários organismos sendo por vezes difícil a procura, organização e comparação da informação. Este modelo poderá ser direcionado para locais onde manipulam e armazenam este tipo de ficheiros, de forma a facilitar a procura e a transferência da informação por parte de utilizadores (bioinformáticos, investigadores, académicos, etc.)

## Planeamento

Como pensamos/solucionamos o problema

# Entidades e atributos

# Modelo conceptual

# Modelo lógico e modelo físico

Os modelos lógicos e físicos da base de dados foram criados no *MySQL Workbench* com base no modelo conceptual apresentado no ponto anterior do presente relatório.

Começando pelo modelo lógico, cada entidade do modelo conceptual corresponde a uma tabela no modelo lógico, nas quais as respetivas colunas correspondem aos atributos. Nesta fase de criação de tabelas para o modelo lógico, são estabelecidas algumas restrições a atribuir a cada atributo, como por exemplo: não nulos, chaves primárias, chaves secundárias, etc., bem como o tipo de dados a inserir.

Ao longo da criação dos modelos, a criação de chaves primárias e secundárias é de extrema importância, uma vez que irão permitir identificar inequivocamente, em cada tabela, uma determinada entrada da base de dados, para além de estabelecer relações entre cada entidade.

Em alguns casos, a chave primária pode ser constituída por um conjunto de chaves estrangeiras. Podemos observar um exemplo deste caso na tabela “locus\_reference”, em que o conjunto das chaves estrangeiras “locus\_GI” e “reference\_reference\_id”, é considerado uma chaves primária, o que significa que um determinado Locus contém um determinado grupo de referências identificado pelos atributos “locus\_GI” e “reference\_reference\_id”, para além de estar identificada a ordem em que cada referência surge do ficheiro Genbank (“ref\_order”).

Podemos também observar um exemplo semelhante ao anterior na tabela “locus\_source”.

# *Functions*, *procedures* e *triggers*

# Povoamento das tabelas

# Discussão e análise

# Bibliografia