Tarea 1: Base de Datos MongoDB para Bioinformática

Universidad de Málaga
Asignatura: Estándares de Datos en Bioinformática y Salud

¿Qué hemos hecho?

Hemos diseñado una base de datos MongoDB para almacenar información relacionada con investigación en bioinformática, específicamente sobre experimentos de secuenciación genética y cáncer.

Las 5 colecciones

Hemos creado 5 "cajones" de información (colecciones) que están conectados entre sí:

1. Experiments (Experimentos)

Aquí guardamos información sobre los experimentos de laboratorio: - Qué tipo de experimento es (RNA-seq) - Qué plataforma se usó (Illumina NovaSeq) - La calidad de los resultados - Quién lo realizó

Ejemplo: Un experimento de análisis RNA-seq en células cancerosas, realizado el 15 de enero de 2024.

2. Samples (Muestras)

Información sobre las muestras biológicas analizadas: - De dónde viene la muestra (tejido tumoral) - Datos del paciente - Qué tratamiento recibió - Cómo se procesó la muestra

Ejemplo: Una muestra de tejido tumoral de mama de un paciente en estadio III tratado con quimioterapia.

3. Genes (Genes)

Datos sobre los genes encontrados: - Nombre del gen (BRCA2) - Cuánto se expresa en diferentes tejidos - Si tiene variantes que causan enfermedades - Con qué otros genes interactúa

Ejemplo: El gen BRCA2, que está relacionado con el cáncer de mama hereditario.

4. Researchers (Investigadores)

Información de los científicos: - Nombre y datos de contacto - Universidad donde trabajan - Su grupo de investigación - Proyectos financiados

Ejemplo: Dra. María García, de la Universidad Complutense, especialista en genómica del cáncer.

5. Publications (Publicaciones)

Artículos científicos publicados: - Título y revista - Número de citas - Dónde están los datos (repositorios como GEO) - Impacto en redes sociales

Ejemplo: Un artículo en Nature sobre análisis genómico del cáncer de mama publicado en 2024.

¿Cómo están conectadas?

Las colecciones están relacionadas entre sí como en la vida real:

- Un **investigador** realiza varios **experimentos**
- Cada **experimento** analiza varias **muestras**
- Las **muestras** contienen varios **genes**
- Los **investigadores** publican **artículos** sobre sus hallazgos
- Los **artículos** hablan de **experimentos** y **genes** específicos

Los 4 niveles de profundidad

Cada colección tiene información organizada en 4 niveles, como cajas dentro de cajas:

Ejemplo en Experiments:

- **Nivel 1**: Datos básicos → "Experimento EXP_001 de tipo RNA-seq"
- **Nivel 2**: Metodología → "Usamos la plataforma Illumina NovaSeq"
- **Nivel 3**: Parámetros técnicos → "Longitud de lectura: 150 pares de bases"
- **Nivel 4**: Control de calidad → "92.5% de bases con alta calidad"

Ejemplos de uso

Con esta base de datos podemos: - Buscar todos los experimentos completados - Encontrar muestras de pacientes con un diagnóstico específico - Ver qué genes tienen alta expresión en tumores - Consultar las publicaciones de un investigador - Analizar la calidad de los experimentos

Conclusión

Hemos creado una base de datos completa y realista que cumple todos los requisitos: - 5 colecciones (pedían mínimo 3) - 4 niveles de anidamiento (pedían mínimo 3) - Todas las colecciones están interconectadas - Datos realistas del ámbito bioinformático - Esquemas bien documentados

Esta base de datos podría usarse en un proyecto real de investigación en cáncer para organizar experimentos, muestras, datos genéticos y publicaciones de forma eficiente.

Repositorio: https://github.com/Husakaa/data-standards-project