GESTIÓN DE DATOS BIOLÓGICOS USANDO BASES DE DATOS RDF

JACQUELINE ANDREA VILLARROEL VALENZUELA
INGENIERÍA CIVIL EN COMPUTACIÓN

RESUMEN

La gestión de datos biológicos es una tarea ardua y complicada debido a la gran cantidad de información que se encuentra disponible y que es necesario explorar. Protein Data Bank (PDB) es una fuente de datos biológica que contiene información sobre proteínas. Actualmente, es muy común usar sistemas de bases de datos relacionales para gestionar datos PDB, sin embargo esto no resulta ser tan apropiado debido principalmente a la organización con estructura de grafo que presentan las proteínas. Esta memoria se enfoca en modelar los datos PDB usando el modelo de datos RDF. Resource Description Framework (RDF) define una manera estándar para modelar datos con estructura de grafo, los cuales pueden ser consultados usando el lenguaje de consulta SPARQL.

RDF es un modelo de datos estándar que posibilita el intercambio y la reutilización de metadatos estructurados, sin las ambigüedades producidas por la procedencia de los datos desde distintas fuentes. Debido a esto es que los nombres en RDF deben ser globales, refiriéndose a que no se debe escoger un nombre que alguien más haya podido concebir para referirse a algo diferente. El modelo de datos RDF se basa en escribir recursos, los cuales son identificados por Identificadores Uniformes de Recursos (URIs). De manera específica, este proyecto partió con el estudio del formato de archivos PDB, el cual está compuesto por varias secciones. La información disponible en cada sección fue modelada a través de un diagrama entidad relación para su mejor comprensión. Habiendo comprendido el contenido de un archivo PDB, y guiados por las indicaciones de investigadores del área de bioinformática, se seleccionó un fragmento de datos correspondiente a proteínas, aminoácidos y átomos, los cuales fueron modelados usando el modelo de datos RDF.A continuación se implementó una herramienta que permite generar un archivo de datos RDF desde un archivo de datos PDB. Con la finalidad de evaluar la usabilidad de los datos, se diseñaron algunas consultas de prueba usando el lenguaje SPARQL. En términos de experimentos, un conjunto de proteínas fueron almacenadas en un sistema de bases de datos RDF, y luego se ejecutaron las consultas de prueba con la finalidad de medir el tiempo de respuesta. Finalmente,

se implementó una herramienta que provee una interfaz sencilla para ejecutar las consultas de prueba.

La principal contribución de este trabajo es el manejo de datos biológicos, ya que la manipulación y análisis de los datos es menos compleja, comparado con el método tradicional que hoy en da utilizan los bioinformáticos. Esto permite que la búsqueda de patrones estructurales en las proteínas sea más rápido sin la necesidad de procesar los datos de forma manual y a la vez cuenta con un lenguaje de consulta estructurado llamado SPARQL.