

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	III
Índice de Figuras	VI
Índice de Tablas	VIII
Resumen	x
1. Introducción	12
1.1. Descripción del problema	12
1.2. Hipótesis	14
1.3. Objetivos	14
1.3.1. Objetivo general	14
1.3.2. Objetivos específicos	14
1.4. Metodología	14
1.5. Alcances	15
1.6. Contribuciones	16
2. Contexto	17
2.1. Datos biológicos	17
2.1.1. Bioinformática	17
2.1.2. Datos sobre proteínas	19
2.1.3. Fuentes de datos biológicos	21
2.1.4. Protein Data Bank (PDB)	24
2.2. The Resource Description Framework (RDF)	32
2.2.1. Modelo de datos RDF	32
2.2.2. Formatos de codificación de datos RDF	34
2.2.3. Esquema de datos RDF	36

2.2.4. SPARQL	37
2.3. Datos biológicos y RDF	38
2.3.1. PDB/RDF	38
2.3.2. Bio2RDF	40
2.3.3. UniprotKB	41
2.3.4. BioLOD	41
3. Modelado de PDB usando el modelo RDF	43
3.1. Introducción	43
3.2. Diagrama entidad relación	46
3.3. Sección “Título”	51
3.4. Sección “Estructura primaria”	63
3.5. Sección “Compuestos Heterogéneos (Ligandos)”	69
3.6. Sección “Estructura secundaria”	71
3.7. Sección “Anotación Conectividad”	76
3.8. Sección “Cristalografía y transformación de coordenadas”	78
3.9. Sección “Coordenadas”	83
3.10. Sección “Conectividad”	87
3.11. Sección “Estadísticas”	88
3.12. Diagrama entidad relación para PDB	89
4. Instancia de datos RDF	93
4.1. RDF Schema para PDB	93
4.2. Vocabulario RDF Schema para PDB	103
4.3. Instancia de datos para PDB	107
5. Experimentos de transformación de datos	119
5.1. Herramienta de transformación de datos	119
5.2. Caso de estudio	124
5.3. Estadísticas sobre los datos transformados	126
6. Experimentos de consulta de datos	130
6.1. Herramienta de consulta de datos	130
6.2. Consultas de prueba	134
6.2.1. Consultas básicas	134

6.2.2. Puente Disulfuro	139
6.2.3. Zinc Finger	141
6.2.4. Motivo P-loop	143
6.3. Análisis de las consultas de prueba	143
7. Conclusiones	150
Bibliografía	153
Anexos	
A: Ejemplo de transformación de datos PDB a RDF	157
A.1. Archivo PDB	157
A.2. Archivo PDB transformado a RDF	158

ÍNDICE DE FIGURAS

	página
2.1. Estructura de las proteínas.	21
2.2. Estructura de un aminoácido.	22
2.3. Formación de un enlace peptídico.	23
2.4. Página oficial de Protein Data Bank (PDB).	24
2.5. Ejemplo resumido de un archivo PDB.	27
2.6. Pasos para el procesamiento de datos PDB.	28
2.7. Página oficial de PDBsum.	30
2.8. Página oficial de PDB Lite.	31
2.9. Página oficial de Protein Explorer.	32
2.10. Grafo RDF que describe los datos de una persona.	34
2.11. Representación RDF Schema de los datos de una persona.	37
2.12. Interfaz de búsqueda de archivos en RDF.	40
3.1. Ejemplo de diagrama entidad relación.	49
3.2. Modelo entidad relación de la bibliografía de una proteína.	62
3.3. Modelo entidad relación de la estructura primaria de una proteína.	68
3.4. Modelo entidad relación del registro Helix de una proteína.	73
3.5. Modelo entidad relación del registro SHEET de una proteína.	75
3.6. Modelo entidad relación de la sección cristalografía y transformación de coordenadas.	82
3.7. Modelo entidad relación de la sección contabilidad de una proteína.	89
3.8. Diagrama entidad relación de los datos PDB de interés.	90
4.1. Modelo RDF Schema de la estructura de una proteína.	104
4.2. Modelo RDF Schema de la distancia entre aminoácidos.	107
4.3. Modelo RDF Schema de la distancia entre átomos	107
5.1. Estructura de datos de BioJava.	121
6.1. Interfaz de consulta de datos PDB-RDF.	131
6.2. Panel que muestra la expresión SPARQL de una consulta.	132
6.3. Manual de ayuda de la aplicación.	134

6.4. Resultados al ejecutar una consulta.	135
6.5. Formación del enlace disulfuro.	140
6.6. Estructura Zinc Finger.	142

ÍNDICE DE TABLAS

	página
2.1. Listado de los aminoácidos estándar	22
3.1. Orden de los registros en un archivo PDB.	48
3.2. Formato del registro HEADER en el archivo PDB.	52
3.3. Formato del registro OBLSTE de una proteína.	53
3.4. Formato del registro TITLE de una proteína.	53
3.5. Formato del registro SPLIT de una proteína.	54
3.6. Formato del registro CAVEAT de una proteína.	54
3.7. Formato del registro COMPND de una proteína.	55
3.8. Formato del registro SOURCE de una proteína.	56
3.9. Formato del registro KEYWDS de una proteína.	56
3.10. Formato del registro EXPDTA de una proteína.	57
3.11. Formato del registro NUMMDL de una proteína.	57
3.12. Formato del registro MDLTYP de una proteína.	58
3.13. Formato del registro AUTHOR de una proteína.	58
3.14. Formato del registro REVDAT de una proteína.	59
3.15. Formato del registro SPRSDE de una proteína.	60
3.16. Formato del registro JRNL de una proteína.	60
3.17. Formato del registro REMARKS de una proteína.	61
3.18. Formato del registro DBREF de una proteína.	64
3.19. Formato del registro DBREF1 de una proteína.	65
3.20. Formato del registro DBREF2 de una proteína.	65
3.21. Formato del registro SEQADV de una proteína.	66
3.22. Formato del registro SEQRES de una proteína.	66
3.23. Formato del registro MODRES de una proteína.	67
3.24. Formato del registro HET de una proteína.	69
3.25. Formato del registro HETNAM de una proteína.	70
3.26. Formato del registro HETSYN de una proteína.	70
3.27. Formato del registro FORMUL de una proteína.	71
3.28. Formato del registro HELIX de una proteína.	72
3.29. Formato del registro SHEET de una proteína.	74

3.30. Formato del registro SSBOND de una proteína.	76
3.31. Formato del registro LINK de una proteína.	77
3.32. Formato del registro CISPEP de una proteína.	78
3.33. Formato del registro CRYST1 de una proteína.	79
3.34. Formato del registro ORIGXn de una proteína.	80
3.35. Formato del registro SCALEn de una proteína.	80
3.36. Formato del registro MTRIXn de una proteína.	81
3.37. Formato del registro MODEL de una proteína.	83
3.38. Formato del registro ATOM de una proteína.	84
3.39. Formato del registro ANISOU de una proteína.	85
3.40. Formato del registro TER de una proteína.	85
3.41. Formato del registro HETATM de una proteína.	86
3.42. Formato del registro CONECT de una proteína.	87
3.43. Formato del registro MASTER de una proteína.	88
5.1. Distribución de los datos en los archivos PDB.	127
5.2. Tamaño de archivos PDB al ser transformados a RDF.	129
6.1. Tiempo de ejecución al buscar la proteína “1TWA”.	144
6.2. Tiempo de ejecución al buscar los aminos concadenados de una proteína.	145
6.3. Tiempo de ejecución al buscar los aminos de una proteína.	145
6.4. Tiempo de ejecución al buscar los aminos ordenados de una proteína.	146
6.5. Tiempo de ejecución al buscar la cadena de una proteína.	146
6.6. Tiempo de ejecución al buscar un amino.	146
6.7. Tiempo de ejecución al buscar dos aminos.	147
6.8. Tiempo de ejecución al buscar un átomo en dos aminos.	147
6.9. Tiempo de ejecución al buscar dos aminos a una cierta distancia.	147
6.10. Tiempo de ejecución al buscar un átomo en dos aminos a una cierta distancia.	148
6.11. Tiempo de ejecución al buscar un enlace disulfuro.	148
6.12. Tiempo de ejecución al buscar un posible Zinc Finger.	149