

UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

Cat-p-Data - Uma ferramenta para a manipulação de dados proteicos

Vinicius Rosa Seus

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande, como requisito para a obtenção do grau de Mestre em Engenharia de Computação

Orientadora: Prof^a. Dr^a. Karina Machado
Co-orientador: Prof. Dr. Adriano Werhli

Rio Grande, abril de 2016

Ficha catalográfica

S496c Seus, Vinicius Rosa.
Cat-p-Data - uma ferramenta para manipulação de dados proteicos /
Vinicius Rosa Seus. – 2016.
77 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande –
FURG, Programa de Pós-graduação em Engenharia de Computação,
Rio Grande/RS, 2016.

Orientadora: Dr^a. Karina Machado.

Coorientador: Dr. Adriano Werhli.

1. Ferramentas automatizadas 2. Dados proteicos 3. Alinhamento de
sequência 4. Alinhamento de estrutura 5. PDB I. Machado, Karina
II. Werhli, Adriano III. Título.

CDU 004.62

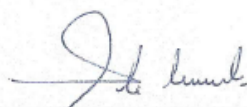
UNIVERSIDADE FEDERAL DO RIO GRANDE
Centro de Ciências Computacionais
Programa da Pós-Graduação em Computação
Curso de Mestrado em Engenharia de Computação

DISSERTAÇÃO DE MESTRADO

**Cat-p-Data - Uma ferramenta para a manipulação de
dados proteicos**

Vinícius Rosa Seus

Banca examinadora:



Profa. Dra. Raquel C. de Melo Minardi (UFMG)



Prof. Dr. Cleo Billa (FURG)



Profa. Dra. Karina dos Santos Machado (FURG)

SUMÁRIO

LISTA DE FIGURAS	6
RESUMO	8
ABSTRACT	9
LISTA DE ABREVIATURAS E SIGLAS	10
1 INTRODUÇÃO	11
1.1 Objetivo geral	12
1.2 Objetivos específicos	13
1.3 Justificativa	13
1.4 Organização do Texto	14
2 FUNDAMENTAÇÃO TEÓRICA	15
2.1 Proteínas e Aminoácidos	15
2.1.1 Mutações pontuais	16
2.2 Bioinformática	18
2.3 Docagem Molecular	19
2.4 Triagem Virtual	21
3 MATERIAIS E MÉTODOS	23
3.1 Banco de dados biológicos públicos utilizados	23
3.1.1 PDB (Protein Data Bank)	23
3.1.2 NCBI/Protein	27
3.1.3 Banco de dados ZINC	27
3.1.4 AAindex	29
3.2 Ferramentas utilizadas no trabalho	29
3.2.1 AutoDock Vina	29
3.2.2 Framework para triagem virtual	31
3.2.3 Ligplot	33
3.3 Tecnologias utilizadas no desenvolvimento da ferramenta Cat-p-Data	33
3.3.1 Biopython: freely available Python tools for computational molecular biology and bioinformatics	33
3.3.2 MySQL	35
3.3.3 Pymol	35
3.3.4 Cytoscape	36

4	TRABALHOS RELACIONADOS	37
4.1	Banco de dados FReDD (Supporting Mining Strategies through a Flexible-Receptor Docking Database)	37
4.2	The PDBbind DataBase	38
4.3	The RS-PDB Database	39
4.4	PDBest: a user-friendly platform for manipulating and enhancing protein structures	40
4.5	Pre-calculated protein structure alignments at the RCSB PDB website	41
4.6	ProtDCal: A program to compute general-purpose numerical descriptors for sequences and 3D-structures of proteins	42
4.7	Tabela comparativa	42
5	RESULTADO 1: FERRAMENTA CAT-P-DATA (CUSTOM ANALYSIS TOOL FOR PROTEIN DATA)	44
5.1	Arquitetura e módulos do Cat-p-data	44
5.1.1	Módulos do Cat-p-data	44
5.1.2	Arquitetura do Cat-p-Data	46
5.2	Banco de dados local	47
5.3	Funcionalidades do Cat-p-data.	50
5.3.1	Funcionalidade 1: Aquisição dos dados e criação do projeto	52
5.3.2	Funcionalidade 2: Alinhamento sequencial	52
5.3.3	Funcionalidade 3: Ferramenta de estudo de similaridade	55
5.3.4	Funcionalidade 4: Alinhamento estrutural	55
5.3.5	Funcionalidade 5: Ferramenta de visualização do ligante	57
5.3.6	Funcionalidade 6: Ambiente tridimensional de visualização de proteínas	57
5.3.7	Funcionalidade 7: Módulo de atualização do conjunto de proteínas	58
5.3.8	Funcionalidade 8: Módulo de limpeza do conjunto de proteínas	59
5.3.9	Funcionalidade 9: Módulo de edição de conjunto de proteínas	60
5.3.10	Funcionalidade 10: Módulo de inserção de novas proteínas	61
6	RESULTADO 2: ESTUDO DE CASO β-GLICOSIDASE	63
6.1	β -glicosidases	63
6.2	Estudo de caso	64
7	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	72
	REFERÊNCIAS	74

LISTA DE FIGURAS

Figura 1	Mutação no aminoácido 191 da β -glicosidase(código PDB:1E55) em destaque na cor vermelha.	18
Figura 2	Resultado da docagem molecular com a proteína da β -glicosidase(código PDB:3WH6) como receptor (em verde) e a “glicose” como ligante (em vermelho).	21
Figura 3	Exemplo do processo de Triagem virtual com N ligantes candidatos, para cada candidato um experimento de docagem molecular é executado.	22
Figura 4	Cabeçalho de um arquivo no formato PDB	24
Figura 5	Formato do registro ATOM de um arquivo PDB.	26
Figura 6	Exemplo do campo ATOM na molécula de código PDB:2BGA	27
Figura 7	Exemplo de arquivo no formato FASTA (sequência primária da proteína de código PDB:2BGA)	27
Figura 8	Exemplo de busca de ligante na base de dados ZINC (aquisição do ligante glicose).	28
Figura 9	Exemplo do formato MOL2 extraído da base de dados ZINC. Parte do arquivo que descreve a estrutura do ligante glicose no formato MOL2.	28
Figura 10	Banco de dados AAindex no formato txt	30
Figura 11	Resultado de um experimento de docagem molecular com o <i>Auto-Dock Vina</i>	31
Figura 12	Área de seleção do receptor, ligante e pasta de resultados.	32
Figura 13	Área de seleção de caixa.	32
Figura 14	Formato <i>Postscript</i> do Ligplot. Exemplo de interação da proteína da β -glucosidase de código PDB:1GAM com o ligante “glicose”.	34
Figura 15	Exemplo de uso da biblioteca <i>Biopython</i>	35
Figura 16	Esquema do banco de dados FReDD	38
Figura 17	Esquema utilizado para a identificação e classificação dos complexos proteína-ligante.	39
Figura 18	Arquitetura da ferramenta PDBest.	41
Figura 19	Módulos do sistema	45
Figura 20	Arquitetura do Cat-p-data	46
Figura 21	Modelagem ER do banco de dados local para o armazenamento dos dados das proteínas.	48
Figura 22	Campos <i>code</i> e <i>description</i> da tabela <i>properties</i> do BD AAIndex após a conversão para o modelo relacional da base de dados proposta.	48

Figura 23	Continuação da tabela <i>properties</i> do banco AAindex.	49
Figura 24	Interface da tela de projetos da ferramenta <i>Cat-p-Data</i> apresentando algumas funcionalidades.	53
Figura 25	Alinhamento sequencial entre duas estruturas primárias. Destaque na cor cinza para o maior segmento de aminoácidos encontrado entre as duas sequências.	55
Figura 26	Exemplo de alinhamento estrutural com algumas proteínas do conjunto da β -glicosidase.	56
Figura 27	Exemplo do uso da ferramenta de visualização do ligante. Proteína de código PDB 1BGG possui dois ligantes GCO (glicose) nas cadeias C e D de sua estrutura.	57
Figura 28	Proteínas do conjunto da β -glicosidase sendo visualizadas no ambiente tridimensional através da nossa ferramenta.	58
Figura 29	Interface do módulo de atualização utilizando como exemplo novas proteínas da β -glicosidase.	59
Figura 30	Interface do módulo de limpeza do conjunto de proteínas utilizando como exemplo o conjunto de proteínas da β -glicosidase.	60
Figura 31	Exemplo do uso do módulo de edição do conjunto de proteínas.	61
Figura 32	Exemplo do uso do módulo de inserção de novas proteínas em um conjunto de proteínas já existente no <i>Cat-p-Data</i>	61
Figura 33	Parcela do alinhamento sequencial do conjunto de proteínas da β -glicosidase.	64
Figura 34	Grupo de β -glicosidases com 90% ou mais de similaridade.	65
Figura 35	β -glicosidase de código PDB:2CBU.	67
Figura 36	Estrutura molecular dos ligantes glicose e celobiose utilizados no experimento de VS do estudo de caso.	67
Figura 37	Resultado do experimento de VS utilizando a glicose como ligante e as β -glicosidases como moléculas-alvo.	68
Figura 38	Resultado do experimento de VS utilizando a celobiose como ligante e as β -glicosidases como moléculas-alvo	68
Figura 39	Interação molecular entre os resíduos da proteína de código PDB:2CBV e os átomos do ligante glicose.	69
Figura 40	Interação molecular entre os resíduos da proteína de código PDB:2WBG e os átomos do ligante celobiose.	70

RESUMO

SEUS, Vinicius Rosa. **Cat-p-Data - Uma ferramenta para a manipulação de dados proteicos**. 2016. 77 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Bancos de dados públicos como o *PDB (Protein Data Bank)*, *NCBI protein* e *ZINC* possuem uma enorme quantidade de dados armazenados. Com esta grande quantidade de dados, fica evidente que o uso de ferramentas que realizam a aquisição e análise desse conjunto de dados é uma necessidade. Além disso, relacionar dados estruturais de proteínas com dados sequenciais de proteínas e relacionar experimentos de docagem molecular e triagem virtual ainda é um desafio. Dessa forma, é apresentado neste trabalho a ferramenta *Cat-p-Data* (do inglês *Custom Analysis tool for protein data*). A ferramenta *Cat-p-Data* integra em um mesmo ambiente diversas funcionalidades para facilitar a pesquisa de dados proteicos. No final deste trabalho, um estudo de caso com o grupo de proteínas da beta-glicosidase é realizado para validar a ferramenta.

Palavras-chave: Ferramentas automatizadas, Dados proteicos, Alinhamento de sequência, Alinhamento de estrutura, PDB.

ABSTRACT

SEUS, Vinicius Rosa. **Cat-p-Data: Custom Analysis Tool for Protein Data**. 2016. 77 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Public databases such as PDB (Protein Data Bank), NCBI protein and ZINC have an enormous amount of stored data. With this large amount of data, it is evident that the use of tools that perform the acquisition and analysis of this data set is a necessity. In addition, relate structural proteins data with sequence proteins data and relate experiments of molecular docking and virtual screening is still a challenge. Thus, it is presented in this paper the Cat-p-Data (Custom Analysis Tool for Protein Data) tool. The tool Cat-p-Data integrates in the same environment several features to facilitate the research of protein data. At the end of this work, a case study with the protein group of beta-glucosidase is performed to validate the tool.

Keywords: Automated tools, Protein data, Sequence alignment, Structure alignment, PDB.

LISTA DE ABREVIATURAS E SIGLAS

CAT-P-Data	- <i>Custom Analysis Tool for Protein Data</i>
DNA	- <i>Deoxyribonucleic acid</i>
FEB	- <i>Free Energy of Binding</i>
FReDD	- <i>Flexible Receptor Docking Database</i>
FTP	- <i>File Transfer Protocol</i>
HTS	- <i>Throughput Virtual Screening</i>
NCBI	- <i>National Center for Biotechnology Information</i>
PDB	- <i>Protein Data Bank</i>
SQL	- <i>Structured Query Language</i>
RNA	- <i>Ribonucleic acid</i>
RS-PDB	- <i>Rich Structure PDB</i>
VMD	- <i>Visual Molecular Dynamics</i>
VS	- <i>Virtual Screening</i>

1 INTRODUÇÃO

Com o grande crescimento de dados proteicos armazenados nas bases de dados disponíveis na *web*, surge a necessidade da criação de diversas aplicações computacionais que auxiliem os pesquisadores na descoberta de conhecimento a partir desses dados. *Protein Data Bank* (PDB) (BERMAN et al., 2000) é um banco de dados público disponível na *web* que contém mais 100 mil estruturas macromoleculares biológicas e disponibiliza a busca e aquisição de dados proteicos através de uma interface *web* no site ¹. Para aquisição desses dados, o usuário realiza uma busca por uma proteína de interesse e, a seguir, pode obter um ou mais arquivos de proteínas, que são armazenados em um diretório no seu computador. A partir desta aquisição de dados, muitas vezes é necessário que o pesquisador analise esse conjunto de estruturas de forma a encontrar dados semelhantes e comparar esses resultados. Essa análise sem um alinhamento sequencial, alinhamento estrutural, ferramenta de visualização de estruturas, aumenta o nível de dificuldade para o pesquisador realizar a sua pesquisa.

O banco de dados do NCBI protein (do inglês, National Center for Biotechnology Information) é uma coleção de dados de sequências de diversas fontes, incluindo traduções de regiões codificadas do GenBank, RefSeq e TPA, bem como registros do SwissProt, PIR, PRF e PDB. As sequências de proteínas são determinantes fundamentais da estrutura e funções biológicas.

Assim, com esta grande quantidade de estruturas de proteínas disponíveis no *PDB* e a grande quantidade de sequências de proteínas no NCBI Protein, além de outros servidores mundiais que armazenam informações diversas de estruturas macromoleculares, torna-se evidente que o uso de ferramentas para a aquisição e análise de conjuntos específicos de macromoléculas biológicas é uma necessidade que visa auxiliar em pesquisas que usam dados proteicos.

Assim, neste trabalho, é proposto o desenvolvimento de uma ferramenta para a aquisição, armazenamento e análises de conjuntos específicos de estruturas de proteínas obtidas a partir do *PDB* em um ambiente que integra diferentes funcionalidades. A ferramenta *Cat-p-Data* é executada no ambiente *desktop* permitindo que o usuário realize

¹<http://www.rcsb.org>

a aquisição das estruturas moleculares a partir do *web-service RESTFUL* fornecido pelo servidor do próprio *PDB*. Após a aquisição de um conjunto de arquivos de estruturas de proteínas (arquivo *PDB*) interessantes, relacionados a um determinado projeto de pesquisa de tema de estudo, o usuário pode manipular esses dados proteicos em um ambiente *off-line* através de um banco de dados local pertencente a ferramenta que armazena informações sobre as características das estruturas, por exemplo, ligantes, mutações, modelos, cadeias, resíduos, átomos, sequência de amino-ácidos (estrutura 1D) e resultados de docagem molecular. Os arquivos de estruturas de proteínas são armazenados localmente no computador do usuário e podem ser utilizados, por exemplo, para simulações de docagem molecular e alinhamento de estruturas primárias e de estruturas terciárias.

Assim, tendo um conjunto de proteínas de interesse disponível e utilizando a ferramenta proposta neste trabalho, o usuário pode realizar análises relacionadas à alinhamentos e visualização das proteínas através de um ambiente gráfico, o qual é integrado na ferramenta. Outras características são relacionadas à exportação de resultados de alinhamentos de sequência primária no formato *CSV (comma separated value)* ou exportação de sequências que possuem uma identidade semelhante em um formato facilmente carregado em ferramentas gráficas. Estes alinhamentos permitem ao usuário visualizar quais proteínas são semelhantes e quais não são em relação a estruturas primária.

Utilizando a ferramenta *Cat-p-Data* proposta neste trabalho, um estudo de caso foi realizado. O estudo de caso tem foco na investigação da sequência de aminoácidos da enzima β -glicosidase. Com a realização dessa investigação, o objetivo foi a caracterização dessa proteína em relação a estrutura primária, terciária e mutações, além de estudos de docagem molecular para verificar a interação com os ligantes glicose e celobiose. O principal foco no estudo de caso era sobre a conclusão sobre quais aminoácidos presentes na estrutura da β -glicosidase tem mais contato com a glicose e celobiose. Para isso, inicialmente uma análise das estruturas da β -glicosidase depositadas no Protein Data Bank (*PDB*) para o entendimento das mutações já cristalografadas, β -glicosidases termoestáveis, ligantes que aparecem interagindo com a proteína, entre outras características, foi necessário. Para a realização desta análise, explorou-se esse conjunto de proteínas da β -glicosidase utilizando a ferramenta desenvolvida neste trabalho. Essa análise visa identificar similaridades das estruturas primárias através de alinhamento sequencial e preparação das estruturas terciárias das proteínas para realização de docagem molecular com os ligantes da glicose e celobiose, entre outras análises.

1.1 Objetivo geral

O objetivo geral deste trabalho consiste na implementação de uma ferramenta para aquisição, armazenamento e análise de conjuntos específicos de estruturas de proteínas obtidos do *PDB*. Com relação à análise de dados, os módulos implementados são os se-

guintes: módulo de alinhamento sequencial de proteínas, módulo de alinhamento estrutural, módulo de visualização de proteínas e ligantes, dentre outros necessários para a facilitar pesquisas que tratam uma quantidade significativa de dados proteicos. A ferramenta implementada neste trabalho é de uso genérico, ou seja, além de ter sido utilizada para o conjunto de proteínas da β -glicosidase, a ferramenta possibilita a aquisição e criação de projetos de quaisquer conjuntos proteicos encontrados no PDB.

1.2 Objetivos específicos

Os objetivos específicos no desenvolvimento desse trabalho são os seguintes:

- Integrar análise de sequências e estruturas de forma a identificar características em comum entre um conjunto de proteínas relacionadas;
- Selecionar conjuntos de estruturas de proteínas para a realização de experimentos de docagem molecular e triagem virtual;
- Criação de projetos específicos de forma local, ou seja, o usuário pode a qualquer momento trabalhar com diferentes grupos de moléculas em um ambiente sem a necessidade de conexão com a internet;
- Realização de um estudo de caso com a enzima da β -glicosidase para a caracterização desta proteína.

1.3 Justificativa

Com o grande crescimento de dados proteicos armazenados nas bases de dados disponíveis na *web*, surge a necessidade da criação de aplicações computacionais que auxiliem os pesquisadores na descoberta de conhecimento a partir desses dados.

A implementação da ferramenta apresentada neste trabalho, é justificada por alguns fatores que dificultam a pesquisa sobre conjuntos de dados estruturais e sequenciais de proteínas. Um destes fatores é a difícil caracterização de um determinado conjunto de proteínas. As proteínas de um conjunto podem ou não compartilhar similaridades na sua estrutura primária, as quais, são relevantes para determinadas pesquisas. Considerando essa dificuldade na caracterização, a ferramenta implementada neste trabalho possui um módulo de alinhamento sequencial para verificar similaridades entre diferentes moléculas. Além disso, o banco de dados local da ferramenta juntamente com outros módulos auxiliam no processo de filtragem das moléculas, armazenando somente proteínas de interesse do pesquisador em cada projeto. Além deste módulo, a ferramenta implementa outros módulos para visualização tridimensional das proteínas, módulo de ligantes, módulo de alinhamento estrutural o qual prepara moléculas para experimentos de triagem virtual, dentre outros módulos para facilitar a pesquisa de conjuntos de proteínas. Após a

aquisição de dados do *PDB*, a ferramenta trabalha em modo *off-line*, sem mais a necessidade de uso de internet.

1.4 Organização do Texto

Este trabalho está organizado da seguinte forma, primeiramente é apresentado o capítulo de fundamentação teórica onde serão discutidos todos os conceitos teóricos fundamentais para o entendimento da ferramenta desenvolvida nesse trabalho. O capítulo seguinte de materiais e métodos apresenta todas as ferramentas e métodos utilizados para a implementação da ferramenta proposta nesse trabalho. O capítulo 4 apresenta os trabalhos relacionados, o qual tem como objetivo apresentar outras ferramentas já existentes para efeito de comparação sobre as funcionalidades de cada uma. Após, o capítulo 5 apresenta a ferramenta Cat-P-Data (Custom Analysis Tool for Protein Data), o qual, apresenta todos os detalhes para a implementação do software. Por último, é apresentado um estudo de caso já utilizando a ferramenta Cat-P-Data.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica necessária para o entendimento da ferramenta desenvolvida nesse trabalho.

2.1 Proteínas e Aminoácidos

Segundo (LESK, 2002), os arquivos de informação em cada organismo existente, o modelo para o desenvolvimento potencial e a atividade de cada indivíduo são relacionadas ao material genético ou DNA, e em alguns vírus ao RNA. Moléculas de DNA são longas, lineares e sua cadeia molecular contém uma mensagem de um alfabeto de 4 letras (A - Adenina, T - Timina, C - Citosina, G - Guanina). Mesmo para microorganismos, esta mensagem é longa, contendo tipicamente 1 milhão de caracteres. Cada DNA implicitamente possui mecanismos para auto-replicação e para tradução de genes em proteínas (LESK, 2002). A informação genética do DNA é de fato uma sequência de letras, onde para cada 3 letras um aminoácido existente na natureza é relacionado. A implementação da informação genética inicialmente ocorre através da síntese de RNA e proteínas, as quais são moléculas responsáveis por grande parte das estruturas e atividades de organismos (LESK, 2002). Para (VERLI, 2014), as proteínas são consideradas polímeros sintetizados pelas células a partir de aminoácidos. São também classificadas como as biomoléculas mais versáteis na natureza, sendo capazes de adotar uma grande possibilidade de arranjos tridimensionais, não encontrada nos demais polímeros.

Segundo (SILVA; SILVA, 2009), aminoácidos são componentes fundamentais dos seres vivos desde o surgimento de vida na Terra. Muitos desses aminoácidos podem ter tido origem extraterrestre, devido ao fato de, eventuais degradações de matérias-orgânicas que entraram na atmosfera terrestre ou através do impacto de cometas ou meteoritos com a superfície do nosso planeta. Ainda, segundo (SILVA; SILVA, 2009), os aminoácidos são pequenas moléculas presentes em seres vivos que caracterizam-se por possuírem um grupo funcional amina (NH_3) e outro carboxílico (COOH) mais um hidrogênio (H) e uma cadeia lateral. Existem algumas moléculas que não possuem estas características, porém são designadas por aminoácidos, como por exemplo, é o caso da taurina. Nos seres vivos,

a importância dos aminoácidos é relevante, podendo ocorrer neles na forma livre ou como constituintes de peptídeos, proteínas, enzimas, polímeros resultantes de condensação que podem incluir outros componentes ou associados a outras moléculas biológicas como é o exemplo da vitamina B (SILVA; SILVA, 2009).

Atualmente, são conhecidos 22 aminoácidos codificados no DNA, os quais são sintetizados por mais de uma via metabólica (SILVA; SILVA, 2009). No entanto, é comum fazer referência ao número de aminoácidos codificado no DNA é 20 (SILVA; SILVA, 2009). Esses 20 aminoácidos existentes no DNA são descritos na Tabela 1 sua abreviação composta de 3 letras e pelo seu nome completo:

As estruturas de proteínas e suas funções são definidas pelas combinações de propriedades físico-químicas e bioquímicas de aminoácidos que ocorrem naturalmente e que são considerados o bloco de construção das proteínas (VERLI, 2014). Segundo (BERG; TYMOCZKO; STRYER, 2002), as estruturas das proteínas podem ser divididas em quatro tipos: estrutura primária, secundária, terciária e quaternária. A estrutura primária apresenta o número de aminoácidos que compõem as moléculas e a sequência ordenada desses resíduos na cadeia polipeptídica (BERG; TYMOCZKO; STRYER, 2002). A estrutura secundária aparece em função das condições físico-químicas do meio onde a proteína está inserida. A estrutura secundária é a disposição regular e repetitiva, a qual a cadeia polipeptídica pode adotar composta por ligações de hidrogênio (BERG; TYMOCZKO; STRYER, 2002). A estrutura terciária é uma estrutura tridimensional completa que é formada através das forças de atração ou repulsão eletrostática, das pontes de hidrogênio e das forças de Van der Waals existentes entre os resíduos de aminoácidos que formam as cadeias (BERG; TYMOCZKO; STRYER, 2002). A estrutura quaternária são compostas por várias moléculas separadas, porém entrelaçadas em estrutura terciária. Se aplica somente a proteínas constituídas por duas ou mais cadeias polipeptídicas, referindo-se a disposição espacial dessas cadeias e as ligações que se estabelecem entre elas como pontes de hidrogênio, atrações eletrostáticas, interações hidrofóbicas (BERG; TYMOCZKO; STRYER, 2002).

2.1.1 Mutações pontuais

Segundo Auerbach (AUERBACH, 2013), a história sobre a pesquisa de mutações pode ser dividida em cinco períodos ao longo do tempo. No primeiro período de 1900 a 1927, os conceitos de mutação e taxas de mutação foram desenvolvidos. Além disso nesse período, foram formuladas as perguntas básicas sobre a natureza da mutação e também as técnicas para as taxas de medição foram trabalhadas. No segundo período, iniciado em 1927, foi descoberta a ação mutagênica de raio-x, a qual proporcionou uma ferramenta adequada para sondar a natureza da mutação. No final deste período, uma nova teoria chamada de teoria alvo foi inventada. De acordo com essa teoria, os genes e as ligações entre eles são alvos de choques de pacotes de energia que são entregues a es-

Tabela 1: Tabela dos aminoácidos existentes na natureza

Nome	Código de 3 letras	Código de 1 letra
Alanina	ala	A
Arginina	arg	R
Asparagina	asn	N
Ácido Aspártico	asp	D
Cisteína	cys	C
Glutamina	gln	Q
Ácido Glutâmico	glu	E
Glicina	gly	G
Histidina	his	H
Isoleucina	ile	I
Leucina	leu	L
Lisina	lys	K
Metionina	met	M
Fenilalanina	phe	F
Prolina	pro	P
Serina	ser	S
Treonina	thr	T
Triptofano	trp	W
Tirosina	tyr	Y
Valina	val	V

tes por meio de radiação ionizante. Quando um gene sofre um choque, ele produz uma mutação pontual (AUERBACH, 2013). O terceiro período iniciou logo antes da segunda guerra mundial, quando surgiu a descoberta de novos agentes mutagênicos químicos, responsáveis pelo surgimento novos problemas. Nesse período surgiram a introdução de micro-organismos adquiridos por meio de análise experimental, principalmente pelo método de ação mutagênica de luz ultravioleta. Segundo Auerbach (AUERBACH, 2013), a maioria das conclusões realizadas no terceiro período exigiam um modelo de química do gene para facilitar a interpretação do mesmo. Porém, neste período nenhum modelo surgiu, caracterizando-o por descobertas no campo experimental e frustrações nas abordagens teóricas. Essa realidade mudou em 1953, quando Watson e Crick baseando-se que o DNA possuía a função comprovada de carregar informações genéticas e considerando os estudos de cristalografia de Wilkins e Rosalind Franklin, propôs o modelo de dupla hélice, o qual é responsável pela fidelidade da replicação do gene e que mantém premissa de que mutações seriam explicáveis em termos de química de ácido-nucleico. Esta descoberta deu início ao quarto período, onde a pesquisa relacionada ao ácido-desoxirribonucleico foi dominada. Esse ácido possui um grande sucesso de execução e continua a fornecer informações sobre a ação dos agentes mutagênicos no DNA (AUERBACH, 2013).

Durante os últimos anos, tem havido uma crescente compreensão de que os mecanismos celulares, como a reparação e a expressão, desempenham um papel importante nos processos de mutação. Além disso, os estudos celulares em mutagênese têm sido con-

siderado como primeiro plano. Assim, o quinto período inicia-se quando, os cientistas compreendem que todos os precedentes devem ser utilizados como base indispensável para a análise de mutação como um processo biológico (AUERBACH, 2013).

Segundo (FEYFANT; SALI; FISER, 2007), a variação entre diferentes tipos de resíduos em um única posição de uma determinada proteína podem gerar consequências importantes com relação à variação genética entre indivíduos.

Essas mutações que ocorrem em um determinado ponto da sequência da proteína podem resultar em uma perda ou alteração da estrutura nativa, que dessa forma pode resultar numa alteração ou perda da função da proteína. Além destas variações naturais entre os indivíduos, pesquisadores da área introduzem um único novo resíduo no lugar de outro em um determinado ponto através da técnica chamada de substituições por mutagênese dirigida com o objetivo de estudar as características estruturais e funcionais de proteínas (FEYFANT; SALI; FISER, 2007).

A Figura 1 apresenta um exemplo de mutação pontual na proteína da β -glicosidase nas cadeias A e B da molécula. A mutação no aminoácido 191 aparece em destaque na Figura.

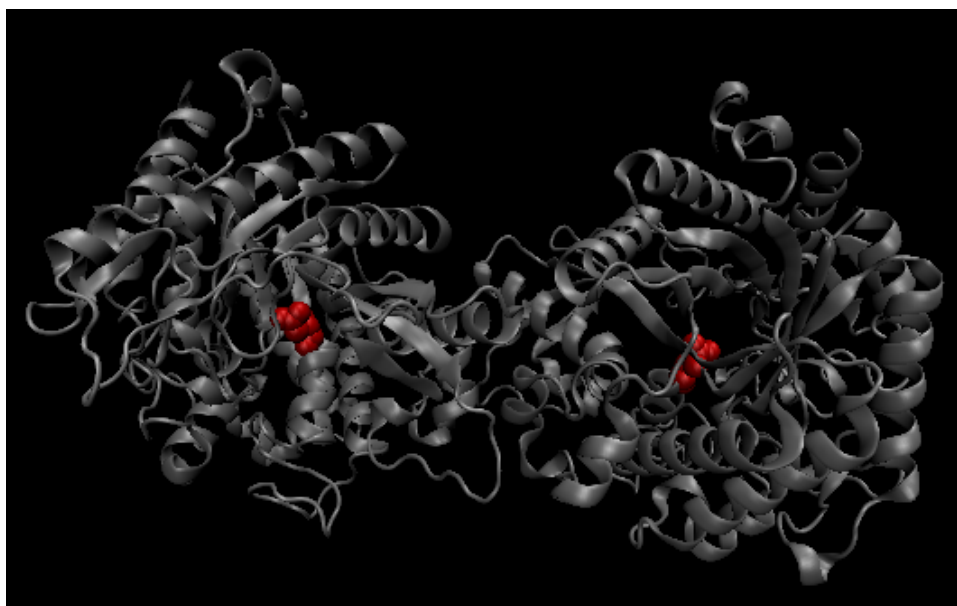


Figura 1: Mutação no aminoácido 191 da β -glicosidase(código PDB:1E55) em destaque na cor vermelha.

Fonte: Elaborado pelo autor utilizando o *software VMD*.

2.2 Bioinformática

A Bioinformática é, segundo (ARNDT, 2008), o uso de tecnologia de informação para a administração de informações biológicas, aplicando técnicas das áreas de computação, estatística e matemática para a resolução de problemas na biologia molecular (LUS-

COMBE; GREENBAUM; GERSTEIN, 2001). Para (VERLI, 2014), a bioinformática pode ser separada em duas vertentes: a bioinformática tradicional, que aborda principalmente problemas relacionados a sequências de nucleotídeos e aminoácidos e pela bioinformática estrutural, que aborda questões biológicas de um ponto de vista tridimensional, envolvendo a maior parte das técnicas compreendidas pela química computacional ou modelagem molecular.

Segundo (OJO; OMABE, 2011) o principal objetivo das pesquisas biológicas é produzir conhecimento que permitirá aproximações preditivas para toda a vida científica. Nesse contexto, a Bioinformática tem importância especial, pois possibilita, através do gerenciamento e da análise de dados em larga escala, uma representação completa da célula e do organismo para sistemas computacionais de predição de alta complexidade (OJO; OMABE, 2011).

Arndt (ARNDT, 2008) afirma que a interdisciplinaridade dessa linha de pesquisa está revolucionando a Biologia. Pois, com o desenvolvimento de softwares e ferramentas virtuais, a visualização das interações moleculares é facilitada, permitindo aos pesquisadores a possibilidade de executar grandes quantidades de testes de predição e aperfeiçoando o entendimento dessas interações. Segundo o autor, uma ênfase é dada aos aspectos visuais das ferramentas, principalmente no que diz respeito à visualização sequencial de dados e a visualização de estruturas proteicas (ARNDT, 2008). Outro aspecto importante que torna a bioinformática necessária, segundo (LESK, 2002), é a larga quantidade de dados que existem para serem processados e analisados.

Cohen (COHEN, 2004) acredita que biólogos computacionais inspiraram-se na biologia para a criação e desenvolvimento de algoritmos importantes que hoje são muito utilizados por biólogos. Biólogos computacionais justificam seus trabalhos, testando algoritmos, estimativas complexas e outras teorias relacionadas às ciências computacionais (COHEN, 2004).

Por essa razão, percebe-se a necessidade de investimentos em pesquisas em Bioinformática, o que propiciará o avanço dos estudos biológicos, permitindo rapidez aos experimentos, diminuindo custos através da utilização de ferramentas para testes *in-silico*, aperfeiçoando os níveis de proximidade desses testes com a realidade, dando viabilidade às pesquisas e aumentando a geração de conhecimento para essa área da ciência (OJO; OMABE, 2011).

2.3 Docagem Molecular

A investigação farmacêutica tem incorporado com sucesso uma grande variedade de métodos de modelagem molecular, dentro de uma variedade de programas relacionados a área de pesquisa de descoberta de fármacos, para investigar sistemas químicos e biológicos complexos (FERREIRA et al., 2015). O conjunto integrado de técnicas com-

putacionais e experimentais tem sido de grande valia na identificação e desenvolvimento de novos compostos promissores (FERREIRA et al., 2015).

Métodos de docagem molecular (do inglês, *Molecular Docking*), são amplamente utilizados no *design* de modernas drogas. Esse método explora as conformações que o ligante pode adotar dentro dos possíveis locais de ligação da molécula alvo. Esta abordagem calcula também a energia livre de ligação (do inglês FEB - *Free Energy of Binding*) através da avaliação de fenômenos críticos envolvidos no processo de reconhecimento intermolecular receptor-ligante (FERREIRA et al., 2015). Atualmente, tem-se uma variedade significativa de algoritmos de docagem molecular disponíveis. Com isso, a compreensão das vantagens e limitações de cada método existente é de fundamental importância para o desenvolvimento de estratégias eficazes e a geração de resultados relevantes (FERREIRA et al., 2015).

Segundo Morris (MORRIS et al., 2009) e Meng (MENG; ZHANG; CUI, 2011), a docagem molecular é uma técnica computacional que é usada para a determinação aproximada da energia de interação entre uma macromolécula receptora e um ligante candidato a inibidor e que também visa encontrar a melhor conformação do ligante. Para (VERLI, 2014), a técnica de docagem molecular dedicada a previsão do modo de ligação e dos detalhes do reconhecimento molecular é dividida em duas classes de métodos distintos, as quais são receptor-ligante e receptor-proteína. Uma das técnicas tem o objetivo principal de descobrir a melhor energia de interação entre proteína-ligante e a outra abordagem busca descobrir o melhor modo de ligação entre proteína-proteína.

A técnica de docagem molecular depende de dois métodos: o primeiro consiste em um campo de força para estimar a energia livre da ligação do complexo e o segundo visa explorar o espaço conformacional disponível para o ligante e a molécula alvo. Na maioria das vezes, é necessário inserir aproximações tanto no método de campo de força como no método de espaço conformacional para que se consiga um processo de docagem com um custo computacional razoável (COSCONATI et al., 2010). Essa abordagem computacional tem como objetivo principal identificar na estrutura da molécula receptora quais são as coordenadas onde a molécula ligante (rígida em sua conformação inicial ou flexível, onde diferentes ângulos de rotação são analisados para o ligante) irá acoplar obtendo um bom resultado em relação à energia livre de ligação (FEB). A partir desse processo, descobrem-se quais são as pequenas moléculas (ligantes) que serão as candidatas a potenciais fármacos (COSCONATI et al., 2010). A Figura 2 apresenta um processo de docagem molecular. O ligante (candidato a fármaco) é representado na cor vermelha, enquanto que a molécula receptora é representada na cor verde (figura desenvolvida utilizando o software Pymol). Dessa forma, se a energia livre de ligação entre o ligante e o receptor for um bom resultado (quanto mais negativa, melhor a energia de ligação), esse ligante se torna um candidato a fármaco.

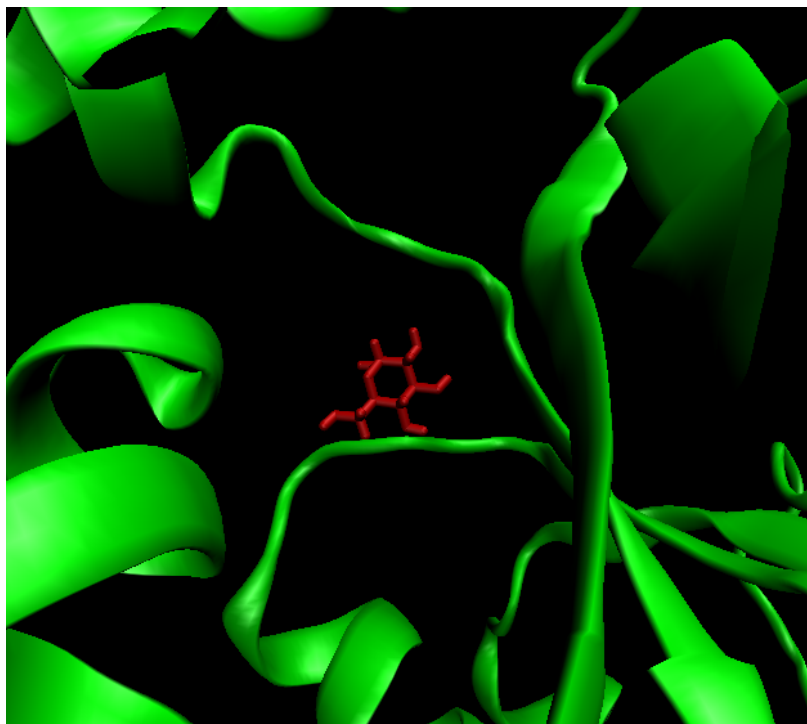


Figura 2: Resultado da docagem molecular com a proteína da β -glicosidase(código PDB:3WH6) como receptor (em verde) e a “glicose” como ligante (em vermelho).

Fonte: Elaborado pelo autor utilizando o *software* VMD.

2.4 Triagem Virtual

A análise *in-silico* de um conjunto de ligantes (candidatos a inibidores) é definida com uma técnica computacional de Triagem Virtual (CAVASSOTO; ORRY, 2007), do inglês *Virtual Screening* (VS). Esse processo testa a interação de proteína-alvo com um conjunto de candidatos a fármacos através de simulações de docagem molecular. Dessa forma, o objetivo consiste em identificar subgrupos de compostos que podem ser adquiridos ou sintetizados e testados experimentalmente, tornando o processo de desenvolvimento de um novo fármaco mais eficaz (KUNTZ, 1992).

A identificação de novas moléculas-alvo e a caracterização de sua estrutura tridimensional vem aumentando em paralelo ao desenvolvimento e avanço dos métodos de Triagem Virtual de alto desempenho HTS (do inglês, High-Throughput Virtual Screening) (MCINNES, 2007). Atualmente, a maioria dos candidatos a fármacos a ser desenvolvidos foram descobertos através do uso de métodos HTS, assim como, algoritmos específicos para fármacos com o objetivo na descoberta de novos medicamentos (MCINNES, 2007). A otimização de métodos de docagem de alto desempenho tem apresentado melhorias na reprodução dos dados experimentais e nas taxas de acerto obtidas. Paralelamente aos métodos de triagem virtual, são estudados também a identificação, design e manipulação de bibliotecas de pequenas moléculas candidatas à fármacos (MCINNES, 2007).

Para Zhao (ZHAO et al., 2009), triagem virtual baseada em estrutura (do inglês,

Structure-based Virtual Screening), consiste no processo de docagem molecular de modelos tridimensionais de drogas candidatas a fármacos em estruturas tridimensionais de moléculas candidatas a receptores e, tornou-se nos últimos anos parte integrante do procedimento para a descoberta de novas drogas. Ainda segundo Zhao (ZHAO et al., 2009), *Virtual Screening* é uma técnica computacional análoga à triagem biológica, a qual é, utilizada para verificar características de um conjunto de moléculas e filtrá-las através de um ou mais procedimentos computacionais. Esses métodos computacionais são mais rápidos e rentáveis do que realizar fisicamente milhares de testes com drogas potenciais ou à base de células. Esses métodos físicos tem sido utilizados pela indústria farmacêutica há décadas (ZHAO et al., 2009). Com relação a comparação entre métodos de HTS e VS, Zhao (ZHAO et al., 2009) sugere que, esses dois métodos devem ser complementares com relação ao seu uso, pois, os resultados de métodos de VS comparados com o HTS apresentam muitas variações.

A Figura 3 mostra um exemplo do processo de triagem virtual. Suponha um receptor-alvo representada pela cor verde na figura 3. O processo de testar diferentes ligantes (representandos na figura 3 pelas cores vermelha e azul) nesta molécula-alvo consiste na docagem molecular de cada ligante.

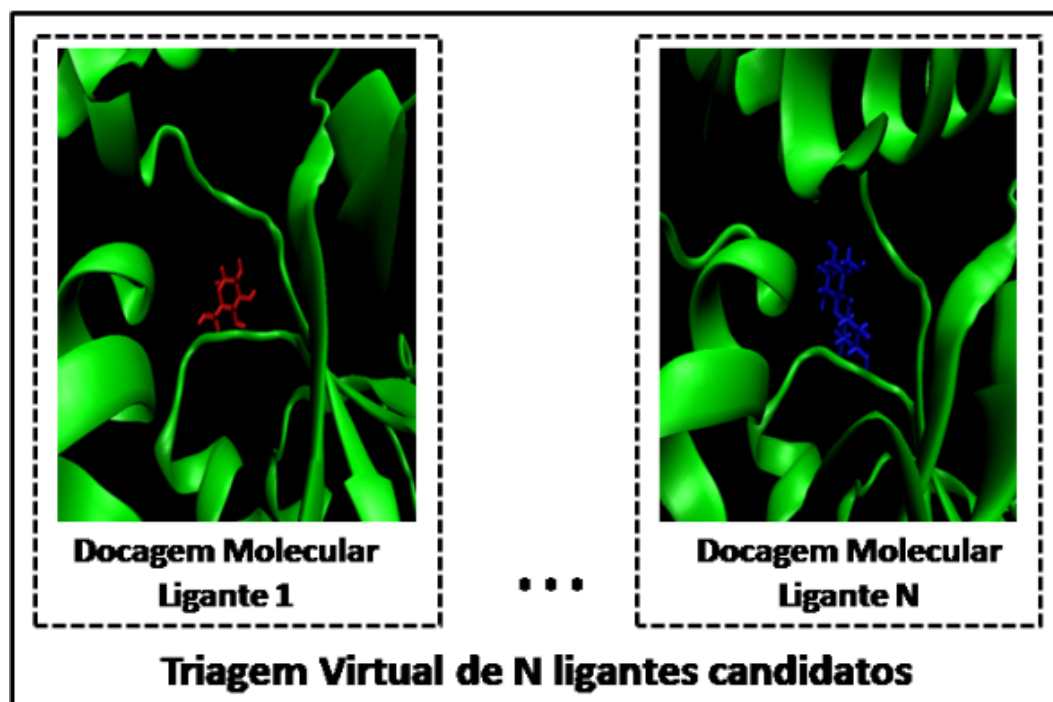


Figura 3: Exemplo do processo de Triagem virtual com N ligantes candidatos, para cada candidato um experimento de docagem molecular é executado.

Fonte: Elaborado pelo autor utilizando o *software VMD*.

3 MATERIAIS E MÉTODOS

Neste capítulo, são apresentados os métodos, ferramentas e sistemas gerenciador de bancos de dados utilizados na criação da ferramenta proposta neste trabalho e no estudo de caso com o conjunto de proteínas da β -glicosidase.

Utilizando a ferramenta produzida neste trabalho em um estudo de caso para filtrar um conjunto mais restrito de importantes proteínas do grupo da β -glicosidase, juntamente com experimentos de triagem virtual e docagem molecular, foi possível localizar através de análise de dados, aminoácidos candidatos à mutações pontuais em algumas proteínas do conjunto. Os ligantes utilizados nos experimentos de triagem virtual foram adquiridos da base de dados do ZINC, já as proteínas, foram obtidas a partir da ferramenta desenvolvida neste trabalho, a qual possui conexão com a base de dados do PDB. Para facilitar a configuração desses experimentos de triagem virtual, foi utilizado um *framework* para triagem virtual, o qual utiliza o *AutoDock Vina* para realizar a docagem molecular. A ferramenta *Ligplot*, foi utilizado para analisar as interações proteína-ligante nos experimentos de triagem virtual realizados. O gerenciador de banco de dados *MySQL* discutido nesta seção, foi utilizado para implementar o banco de dados local da ferramenta desenvolvida neste trabalho, onde as proteínas e suas características são armazenadas após a aquisição do PDB. O *BioPython*, biblioteca de bioinformática para *Python*, foi utilizado para facilitar a extração de informações de arquivos de proteínas. Como experimento futuro, após a realização das mutações, pretende-se realizar novamente experimentos de triagem virtual para analisar o novo comportamento das moléculas da β -glicosidase.

3.1 Banco de dados biológicos públicos utilizados

3.1.1 PDB (Protein Data Bank)

Protein Data Bank (PDB) é um banco de dados que foi criado em 1971 no Laboratório Nacional de Brookhaven. Inicialmente era um arquivo para estruturas cristalinas biológicas de macromoléculas. Nos seus primeiros anos armazenou 7 estruturas macromoleculares. Nos anos de 1980, o número de estruturas depositadas começou a aumentar drasticamente. Atualmente é um banco de dados *web* público que possuía, até janeiro de

2016, mais de 100 mil estruturas macromoleculares biológicas depositadas (BERMAN et al., 2007). Os dados contidos no PDB são gerados e enviados por pesquisadores e cientistas de todo o mundo para sites localizados nos Estados Unidos, Europa e Ásia (BERMAN et al., 2007).

Segundo (BERMAN et al., 2000) um arquivo *PDB* é constituído de um conjunto de coordenadas atômicas e outras informações que descrevem proteínas e outras importantes macromoléculas biológicas. Biólogos estruturais utilizam métodos experimentais para determinar a localização de cada átomo em relação a outros átomos na mesma molécula como, por exemplo, cristalografia por difração de raios-X e ressonância magnética nuclear. Esses pesquisadores, então, depositam essas informações, as quais são anotadas e lançadas publicamente no arquivo pelo wwPDB5. Este arquivo é mantido por organizações que funcionam como depósito, executam processamento de dados e são centros de distribuição para os dados no formato *PDB*. Um exemplo do cabeçalho no formato PDB é exemplificado na figura 4 que descreve o arquivo PDB com código 1E4L.

```

HEADER      HYDROLASE                               10-JUL-00   1E4L
TITLE       CRYSTAL STRUCTURE OF THE INACTIVE MUTANT MONOCOT
TITLE       2 (MAIZE ZMGLU1) BETA-GLUCOSIDASE ZM GLU191ASP
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: BETA-GLUCOSIDASE, CHLOROPLASTIC;
COMPND      3 CHAIN: A, B;
COMPND      4 SYNONYM: BETA-D-GLUCOSIDE GLUCOHYDROLASE, CELLOBIASE, GENTIOBIASE;
COMPND      5 EC: 3.2.1.21;
COMPND      6 ENGINEERED: YES;
COMPND      7 MUTATION: YES
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: ZEA MAYS;
SOURCE      3 ORGANISM_COMMON: MAIZE;
SOURCE      4 ORGANISM_TAXID: 4577;
SOURCE      5 STRAIN: CV. MUTIN;
SOURCE      6 TISSUE: COLEOPTILE;
SOURCE      7 ORGANELLE: CHLOROPLAST;
SOURCE      8 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE      9 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE     10 EXPRESSION_SYSTEM_CELL: PLYS S CELLS;
SOURCE     11 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE     12 EXPRESSION_SYSTEM_VECTOR: PET-21A;
SOURCE     13 EXPRESSION_SYSTEM_GENE: GLU1
KEYWDS      GLYCOSIDE HYDROLASE, HYDROLASE, FAMILY 1, RETENTION
KEYWDS      2 OF THE ANOMERIC CONFIGURATION, INACTIVE MUTANT E191D
EXPDTA      X-RAY DIFFRACTION
AUTHOR      M.CZJZEK,M.CICEK,D.R.BEVAN,B.HENRISSAT,A.ESEN
REVDAT      3  24-AUG-11 1E4L   1      HEADER COMPND SOURCE KEYWDS
REVDAT      3  2      REMARK HETSYN FORMUL VERSN
REVDAT      2  24-FEB-09 1E4L   1      VERSN
REVDAT      1  11-DEC-00 1E4L   0
JRNL        AUTH  M.CZJZEK,M.CICEK,V.ZAMBONI,D.R.BEVAN,B.HENRISSAT,
JRNL        AUTH  2 A.ESEN

```

Figura 4: Cabeçalho de um arquivo no formato PDB

O formato PDB possui um cabeçalho padronizado através do uso de palavras reservadas no arquivo. Em cada linha, diferentes palavras reservadas aparecem. Conforme é

mostrado na Figura 4, cada palavra reservada do cabeçalho de um arquivo PDB é listada e discutida a seguir.

- **HEADER** - armazena a descrição da proteína, assim como a classificação da molécula, a data de depósito da proteína no PDB e o código PDB único de determinada proteína;
- **SOURCE** - armazena a origem biológica e/ou química específica de cada molécula;
- **AUTHOR** - armazena os nomes das pessoas responsáveis pelos conteúdos de entrada;
- **OBSLTE** - aparece em entradas que foram removidas da distribuição pública;
- **KEYWDS** - são armazenadas informações referentes às palavras-chave para a busca da molécula;
- **REVDAT** - armazena um histórico das modificações feitas a uma entrada desde seu lançamento;
- **TITLE** - armazena o título dado à pesquisa referente à determinada proteína;
- **EXPDTA** - armazena informações sobre o experimento;
- **SPRSDE** - armazena uma lista de códigos de identificação de entradas que foram obsoletas devido a uma coordenada e removidas do conjunto do PDB;
- **SPLIT** - é usado nos casos em que uma entrada específica compõem parte de um grande complexo macromolecular. A palavra reservada **SPLIT** vai identificar os PDBs que são necessários para reconstituir um complexo completo;
- **NUMMDL** - armazena o número total de modelos em um PDB;
- **JRNL** - armazena a citação primária na literatura que descreve o experimento, que resultou no conjunto de coordenadas depositadas. Outras referências são informadas em **REMARK 1**;
- **CAVEAT** - alerta para erros e problemas não resolvidos no PDB.
- **MDLTYP** - contém anotação adicional pertinente para as coordenadas apresentadas no PDB;
- **REMARKS** - contém detalhes experimentais, anotações, comentários e informações não incluídas em outros registros;

- COMPND - são armazenadas informações referentes ao nome da molécula, quantas e quais cadeias que a molécula possui, sinônimos do nome da molécula, além de indicar se a molécula é uma mutação ou não;
- HET - são armazenadas informações referentes aos resíduos não-convencionais, tais como, grupos prostéticos, inibidores de moléculas de solvente, além de íons para os quais as coordenadas são fornecidas;
- HETATM - armazena átomos de moléculas de água ou átomos apresentados no grupos “HET”.
- TER - indica o fim de uma lista de registros do tipo ATOM/HETATM para uma cadeia;
- ATOM - armazena as coordenadas atômicas para aminoácidos e nucleotídeos padrão. Apresenta também o fator de ocupação e temperatura para cada átomo. Coordenadas químicas não-poliméricas usam o tipo de registro HETATM. O símbolo do elemento químico está sempre presente em cada registro ATOM. A carga do átomo é opcional.

A Figura 5 apresenta o formato padrão do campo ATOM em um arquivo *PDB* e a Figura 6 apresenta o campo ATOM na molécula de código PDB:2BGA.

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

Figura 5: Formato do registro ATOM de um arquivo PDB.

Fonte: Extraído de

“<http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM>”

Para a realização da aquisição de arquivos de proteínas, a ferramenta desenvolvida neste trabalho conecta-se ao *web service RESTful* do PDB. O banco de dados PDB suporta o *web service interface RESTful* para facilitar a aquisição de dados moleculares.

ATOM	1	N	THR	A	2	61.990	84.851	70.908	1.00	32.14	N
ATOM	2	CA	THR	A	2	62.828	85.531	69.934	1.00	20.14	C
ATOM	3	C	THR	A	2	62.799	85.087	68.476	1.00	17.99	C
ATOM	4	O	THR	A	2	62.825	83.895	68.164	1.00	18.99	O
ATOM	5	CB	THR	A	2	64.279	85.578	70.437	1.00	17.25	C
ATOM	6	OG1	THR	A	2	64.266	85.699	71.855	1.00	15.05	O
ATOM	7	CG2	THR	A	2	65.017	86.785	69.894	1.00	15.54	C
ATOM	8	N	ILE	A	3	62.751	86.085	67.597	1.00	13.43	N
ATOM	9	CA	ILE	A	3	62.752	85.907	66.153	1.00	15.97	C
ATOM	10	C	ILE	A	3	64.130	86.302	65.611	1.00	15.81	C
ATOM	11	O	ILE	A	3	64.667	87.352	65.960	1.00	14.99	O
ATOM	12	CB	ILE	A	3	61.640	86.759	65.487	1.00	17.05	C
ATOM	13	CG1	ILE	A	3	60.296	86.031	65.635	1.00	16.08	C

Figura 6: Exemplo do campo ATOM na molécula de código PDB:2BGA

Fonte: Elaborado pelo autor

3.1.2 NCBI/Protein

O NCBI protein (do inglês, National Center for Biotechnology Information)¹ é uma coleção de dados de sequências de diversas fontes, incluindo traduções de regiões codificadas do GenBank, RefSeq e TPA, bem como registros do SwissProt, PIR, PRF e PDB. As sequências de proteínas (estruturas primárias) são determinantes fundamentais da estrutura e funções biológicas.

A seguir, a Figura 7 apresenta um exemplo de arquivo no formato FASTA, o qual armazena a estrutura primária (sequência de aminoácidos) da molécula de código PDB:2BGA.

```
>1BGA:A|PDBID|CHAIN|SEQUENCE
TIFQFPQDFMWGTATAAYQIEGAYQEDGRGLSIWDTFAHTPGKVFNGDNGNVACDSYHRYEEDIRLMKELGIRTYRFSVS
WPRIFPNGDGEVNVQEGLDYYHRVVDLLNDNGIEPFCTLYHWDLPQALQDAGGWGNRRTIQAFVQFAETMFREFHGKIQHW
LTFNEPWCIAFLSNMLGVHAPGLTNLQTAIDVGHHLLVAHGLSVRRFRELGTSGQIGIAPNVSWAVPYSTSEEDKAACAR
TISLHSDWFLQPIYQGSYPQFLVDWFQEGATVPIQDGDMDIIGEPIDMIGINYYSMSVNRFNPEAGFLQSEEINMGFLPV
TDIGWPVESRGLYEVLHYLQKYGNIDIYITENGACINDEVVNGKVQDDRRISYMQQHLVQVHRTIHDGLHVKGYMAWSLL
DNFEWAEGYNMRFGLMIHVDVFRITQVRTPKESYYWYRNVVSNWLETRR
```

Figura 7: Exemplo de arquivo no formato FASTA (sequência primária da proteína de código PDB:2BGA)

Fonte: Elaborado pelo autor

3.1.3 Banco de dados ZINC

ZINC (IRWIN et al., 2012) é uma fonte de dados pública que tem como objetivo principal disponibilizar compostos químicos (ou ligantes) desenvolvidos por empresas farmacêuticas ou universidades para uso em testes *in-silico*. Consiste em um banco de dados com mais de trinta e cinco milhões de moléculas comercialmente disponíveis em representações biológicas relevantes. Essas moléculas podem ser obtidas de uma interface *web* individualmente ou em subconjuntos em um formato *ready-to-dock*, que são arqui-

¹<http://www.ncbi.nlm.nih.gov/protein>

vos que possuem informações sobre as moléculas e estão em um formato pronto para o processo de docagem. A estrutura do site permite a busca nas seguintes formas: estrutura, atividade biológica, propriedade do fornecedor, número do catálogo, nome, entre outras.

A base de dados ZINC foi utilizada neste trabalho para a aquisição das estruturas moleculares dos ligantes utilizados no estudo de caso descrito no capítulo 6. As Figuras 8 e 9 apresentam respectivamente a busca do ligante “glicose” na base de dados do ZINC e a sua representação no formato MOL2. Referente à busca apresentada na Figura 8, o usuário pode adquirir todas os ligantes listados na busca através de duas formas: em uma delas, o usuário pode obter todos os ligantes da busca em um mesmo arquivo e a outra forma é realizar aquisição de ligantes separadamente. Com relação ao formato, a base de dados do ZINC oferece às seguintes possibilidades para aquisição: MOL2, SDF, SMILES e Flexibase.

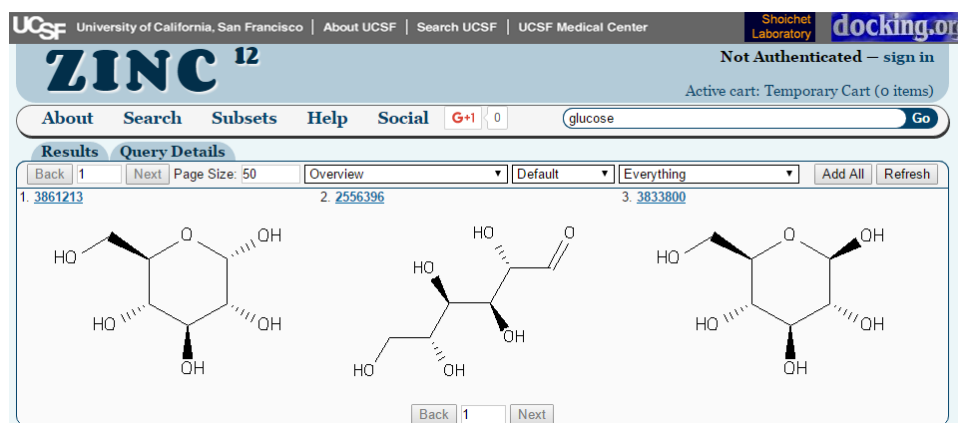


Figura 8: Exemplo de busca de ligante na base de dados ZINC (aquisição do ligante glicose).

Fonte: Extraído de “<http://zinc.docking.org/synonym/glucose>”

```
@<TRIPOS>MOLECULE
ZINC03861213
  24    24    0    0    0
SMALL
USER_CHARGES

@<TRIPOS>ATOM
  1 C1      -0.0187    1.5258    0.0104 C.3    1 <0>    0.0886
  2 C2       0.0021   -0.0041    0.0020 C.3    1 <0>    0.1003
  3 H1       0.5883   -0.3550   -0.8473 H    1 <0>    0.0876
  4 C3      -1.4298   -0.5335   -0.1124 C.3    1 <0>    0.0943
  5 H2      -2.0272   -0.1498    0.7147 H    1 <0>    0.0743
  6 C4      -1.4057   -2.0641   -0.0595 C.3    1 <0>    0.0838
  7 H3      -0.8501   -2.4498   -0.9143 H    1 <0>    0.0838
  8 C5      -0.7236   -2.5086    1.2379 C.3    1 <0>    0.0587
  9 H4      -1.3066   -2.1648    2.0923 H    1 <0>    0.0878
 10 C6       0.6811   -1.9029    1.3002 C.3    1 <0>    0.2260
```

Figura 9: Exemplo do formato MOL2 extraído da base de dados ZINC. Parte do arquivo que descreve a estrutura do ligante glicose no formato MOL2.

Fonte: Elaborado pelo autor

3.1.4 AAindex

Segundo (KAWASHIMA et al., 2008) uma grande variedade de propriedades de aminoácidos foram investigadas por meio de um considerável número de experimentos e estudos teóricos. Dessa forma, foi concluído que cada uma destas propriedades de aminoácidos pode ser representada por um conjunto de 20 valores numéricos e também referenciada por um índice de aminoácido. Assim, foi criado o banco de dados chamado de AAindex (do inglês “amino acid index database”) que consiste em um conjunto de índices numéricos que representam várias propriedades físico-químicas e bioquímicas e pares de aminoácidos. Segundo (KAWASHIMA et al., 2008), AAindex está sendo usado para amplas pesquisas na área de bioinformática, como a investigação sobre sequências de proteínas, predição da localização subcelular de proteínas, a imunogenicidade de ligação de peptídeos da classe MHC I e substituições de coordenadas em vários alinhamentos de sequências de proteínas. Com a necessidade de se utilizar o banco de dados AAindex nesta pesquisa em trabalhos futuros, obteve-se a base de dados através do link (http://www.genome.jp/dbget-bin/www_bfind?aaindex). Essa base de dados é adquirida em um simples arquivo com formato “txt”(Figura 10). A base de dados AAindex faz parte do banco de dados local da ferramenta desenvolvida neste trabalho, porém, será utilizada como um trabalho futuro na busca de características dos aminoácidos.

3.2 Ferramentas utilizadas no trabalho

3.2.1 AutoDock Vina

A ferramenta utilizada para a execução do processo de docagem molecular no *framework* para triagem virtual (seção 3.2.2) é o *software AutoDock Vina*. Trata-se de um programa para docagem molecular e triagem virtual. Segundo Trott e Olson (TROTT; OLSON, 2010) *AutoDock Vina*, possibilita o uso de ligantes flexíveis, assim como receptores que possuem cadeias laterais de sítio de ligação flexíveis. Ainda segundo Trott e Olson (TROTT; OLSON, 2010), o *AutoDock Vina* tem função de *score* projetada para trabalhar com o somatório de todos os pares de átomos que podem mover-se uns aos outros, possuindo bons resultados em relação a sua precisão e, além disso, possibilita o uso de múltiplos núcleos de processamento do computador, o que resulta na execução mais rápida dos experimentos. O *AutoDock Vina* trabalha somente com receptores e ligantes no formato “pdbqt”. Após a realização de uma simulação de docagem molecular são gerados 2 arquivos de saída. Um arquivo é gerado no formato PDBQT, o qual apresenta os modelos e coordenadas do ligante ao final do experimento. Esses modelos representam as tentativas de atracamento do ligante no receptor. O outro arquivo é gerado no formato de log(11), o qual lista os melhores resultados de interação entre receptor-ligante. O arquivo de log da figura 11 apresenta um resultado de docagem molecular, onde o campo

```

H ANDN920101
D alpha-CH chemical shifts (Andersen et al., 1992)
R LIT:1810048b PMID:1575719
A Andersen, N.H., CaQ, B. and Chen, C.
T Peptide/protein structure analysis using the chemical shift index method:
  upfield alpha-CH values reveal dynamic helices and aL sites
J Biochem. and Biophys. Res. Comm. 184, 1008-1014 (1992)
C BUNA790102 0.949
I   A/L   R/K   N/M   D/F   C/P   Q/S   E/T   G/W   H/Y   I/V
    4.35   4.38   4.75   4.76   4.65   4.37   4.29   3.97   4.63   3.95
    4.17   4.36   4.52   4.66   4.44   4.50   4.35   4.70   4.60   3.95
//
H ARG820101
D Hydrophobicity index (Argos et al., 1982)
R LIT:0901079b PMID:7151796
A Argos, P., Rao, J.K.M. and Hargrave, P.A.
T Structural prediction of membrane-bound proteins
J Eur. J. Biochem. 128, 565-575 (1982)
C JOND750101   1.000 SIMZ760101   0.967 GOLD730101   0.936
  TAKK010101   0.906 MEEJ810101   0.891 ROSM880104   0.872
  CIDH920105   0.867 LEVM760106   0.865 CIDH920102   0.862
  MEEJ800102   0.855 MEEJ810102   0.853 ZHOH040101   0.841
  CIDH920103   0.827 PLIV810101   0.820 CIDH920104   0.819
  LEVM760107   0.806 NOZY710101   0.800 GUYH850103  -0.808
  PARJ860101  -0.835 WOLS870101  -0.838 BULH740101  -0.854
I   A/L   R/K   N/M   D/F   C/P   Q/S   E/T   G/W   H/Y   I/V
    0.61   0.60   0.06   0.46   1.07   0.   0.47   0.07   0.61   2.22
    1.53   1.15   1.18   2.02   1.95   0.05   0.05   2.65   1.88   1.32
//
H ARG820102
D Signal sequence helical potential (Argos et al., 1982)
R LIT:0901079b PMID:7151796
A Argos, P., Rao, J.K.M. and Hargrave, P.A.
T Structural prediction of membrane-bound proteins
J Eur. J. Biochem. 128, 565-575 (1982)
C ARG820103   0.961 KYTJ820101   0.803 JURD980101   0.802
I   A/L   R/K   N/M   D/F   C/P   Q/S   E/T   G/W   H/Y   I/V
    1.18   0.20   0.23   0.05   1.89   0.72   0.11   0.49   0.31   1.45
    3.23   0.06   2.67   1.96   0.76   0.97   0.84   0.77   0.39   1.08

```

Figura 10: Banco de dados AAindex no formato txt
Elaborado pelo autor. Adquirido de “<http://www.genome.jp/aaindex>”.

affinity(kcal/mol) representa a FEB.

O *AutoDock Vina* é uma nova versão para o *software* anterior *AutoDock 4*, apresentando melhorias em relação à aceleração na docagem, à precisão das previsões dos modos de ligação e ao cálculo automático dos mapas de rede e *clusters*, segundo Trott e Olson (TROTT; OLSON, 2010).

mode	affinity	dist from best mode	
	(kcal/mol)	rmsd l.b.	rmsd u.b.
1	-6.4	0.000	0.000
2	-5.9	1.212	4.186
3	-5.8	1.488	3.100
4	-5.6	1.910	3.739
5	-5.4	2.219	3.294
6	-5.3	5.914	7.090
7	-5.2	6.219	7.837
8	-5.1	2.748	4.231
9	-5.1	2.743	4.616

Figura 11: Resultado de um experimento de docagem molecular com o *AutoDock Vina*.

Fonte: Elaborado pelo autor. Extraído da saída padrão do *software AutoDock Vina*.

3.2.2 Framework para triagem virtual

Esta seção apresenta o *framework* para triagem virtual a ser utilizado neste trabalho. Este *framework* foi desenvolvido no meu trabalho de conclusão de curso e aprimorado com a ajuda de outros colaboradores. A ferramenta foi publicada no evento Symposium on Applied Computing (SAC) 2016 (SEUS et al., 2016).

A partir de diferentes configurações para um determinado experimento de VS, o usuário pode, ao final da configuração do seu experimento, obter um ou um conjunto de *scripts* em *Python* que, quando executados localmente em sua máquina, realizam o experimento de triagem virtual anteriormente configurado na interface (Figuras 12 e 13), de forma simples e prática (SEUS et al., 2016). O *framework* utiliza o *AutoDock Vina* para executar a docagem molecular. Para a realização de um experimento de docagem molecular, o usuário deve possuir em seu computador a molécula-alvo e um ligante. Além disso, o usuário deve ter conhecimento prévio de uma possível área de docagem na molécula-alvo, para dessa forma informar na interface do *framework* essa informação. Após um experimento de docagem ser realizado, o *AutoDock Vina* gera um arquivo no formato apresentado na figura 11.

Através da interface do *framework* (Figuras 12 e 13) o usuário poderá selecionar um



Virtual Screening Framework

an easy way to prepare virtual screening



Input/Output Area

(Click here for hide/show)

Do you want generate the script for Windows or Linux?

☐ Windows
 ☒ Linux

Do you want execute the same experiment more than once?

☒ Yes
 ☐ No

Flexible Receptor (multiple receptors in the filled path). Mark this option if you have many receptors to consider in the virtual screening.

☐

Select the Receptor Type:

☐ PDB
 ☒ PDBQT
 ☐ PDBQT

Receptor path:

Paste or write here the receptor path followed by receptor name in your computer. Example: C:\script_vs\11IWG

Select the Ligand Type:

☐ PDBQT
 ☒ MOL2(file with only one structure)
 ☐ MOL2(file with multiple structures)

Ligand path:

Paste or write here the ligand path followed by ligand name in your computer. Example: C:\script_vs\mlm18_hydro

☒ Rigid Ligand

Output Folder:

Paste or write here the output path for the results. Example: C:\script_vs\out_vs\

Figura 12: Área de seleção do receptor, ligante e pasta de resultados.
Fonte: Elaborado pelo autor.

Box Area

(Click here for hide/show)

Box Size:

X

Y

Z

Box Center:

X

Y

Z

Do you want variate the center of the box? (Execute the virtual screening considering different coordinates.)

☒ Yes
 ☐ No

Do you want LigPlot results?

☒ Yes
 ☐ No

Ligplot Folder:

Paste or write here the ligplot path. Example: C:\ligplot\exe\

Generate Script Ligplot

Generate Script

Developed by Vinicius Rosa Seus - Karina dos Santos Machado - Jorge H. Gomes

Figura 13: Área de seleção de caixa.
Fonte: Elaborado pelo autor.

ou mais receptores e na área do ligante selecionar um ou mais ligantes. Após esta primeira configuração, o usuário caracteriza a caixa (Figura 13), ou seja, indica na interface do *framework* (área chamada de “Box Area”) a coordenada onde a docagem molecular com o receptor e o ligante acontecerá. Por fim, o usuário clica no botão “Generate Script” e dessa forma realiza o *download* do *script* e executa em seu próprio computador. Essa execução da docagem molecular é realizada com o uso do *software AutoDock Vina*.

O *framework* oferece ao usuário também a possibilidade de inclusão de uma análise com a ferramenta *LigPlot*, a qual analisa a interação de cada átomo do ligante com os aminoácidos do receptor, para isso, o usuário indica na interface do *framework* que deseja resultados do *LipPlot* e clica em “Generate Script Ligplot”. Após a execução dos *scripts*, o usuário pode analisar os resultados da triagem virtual e verificar o novo comportamento da proteína-alvo.

3.2.3 Ligplot

Segundo (WALLACE; LASKOWSKI; THORNTON, 1996) *Ligplot* é um programa que gera automaticamente representações esquemáticas 2D de complexos de proteína-ligante. A saída do programa pode ser colorida ou em preto e branco. Um arquivo de saída do programa no formato *Postscript* fornece uma simples e informativa representação das interações intermoleculares e os seus pontos fortes, incluindo ligações de hidrogênio, interações hidrofóbicas e acessibilidades dos átomos. Ainda segundo (WALLACE; LASKOWSKI; THORNTON, 1996), *Ligplot* é completamente genérico para qualquer tipo de ligante e também pode ser utilizado para visualizar outros tipos de interações em proteínas e ácidos nucleicos. *Ligplot* foi projetado para facilitar a rápida inspeção de muitos complexos enzimáticos, porém é utilizado para outras muitas aplicações.

A Figura 14 apresenta o formato gerado pelo *Ligplot*, apresentando a interação dos resíduos da proteína de código PDB:1GAM com os átomos do ligante “glicose”. Como pode-se analisar na Figura 14, os aminoácidos da proteína Thr 62, ASP 65, LEU 58, entre outros indicados pela linha circular tracejada correspondem aos aminoácidos que, após a docagem molecular estabeleceram contatos do tipo *non-bonded* com o ligante glicose. Os aminoácidos Hys 61 e Arg 428 são aqueles que estabeleceram ligações de hidrogênio.

3.3 Tecnologias utilizadas no desenvolvimento da ferramenta Cat-p-Data

3.3.1 Biopython: freely available Python tools for computational molecular biology and bioinformatics

Segundo (COCK et al., 2009), *Biopython* é uma ferramenta madura, de código livre e que provê bibliotecas em *Python* que ajudam em uma ampla gama de problemas encontrados na bioinformática. *Biopython* possui módulos para leitura e escrita de arqui-

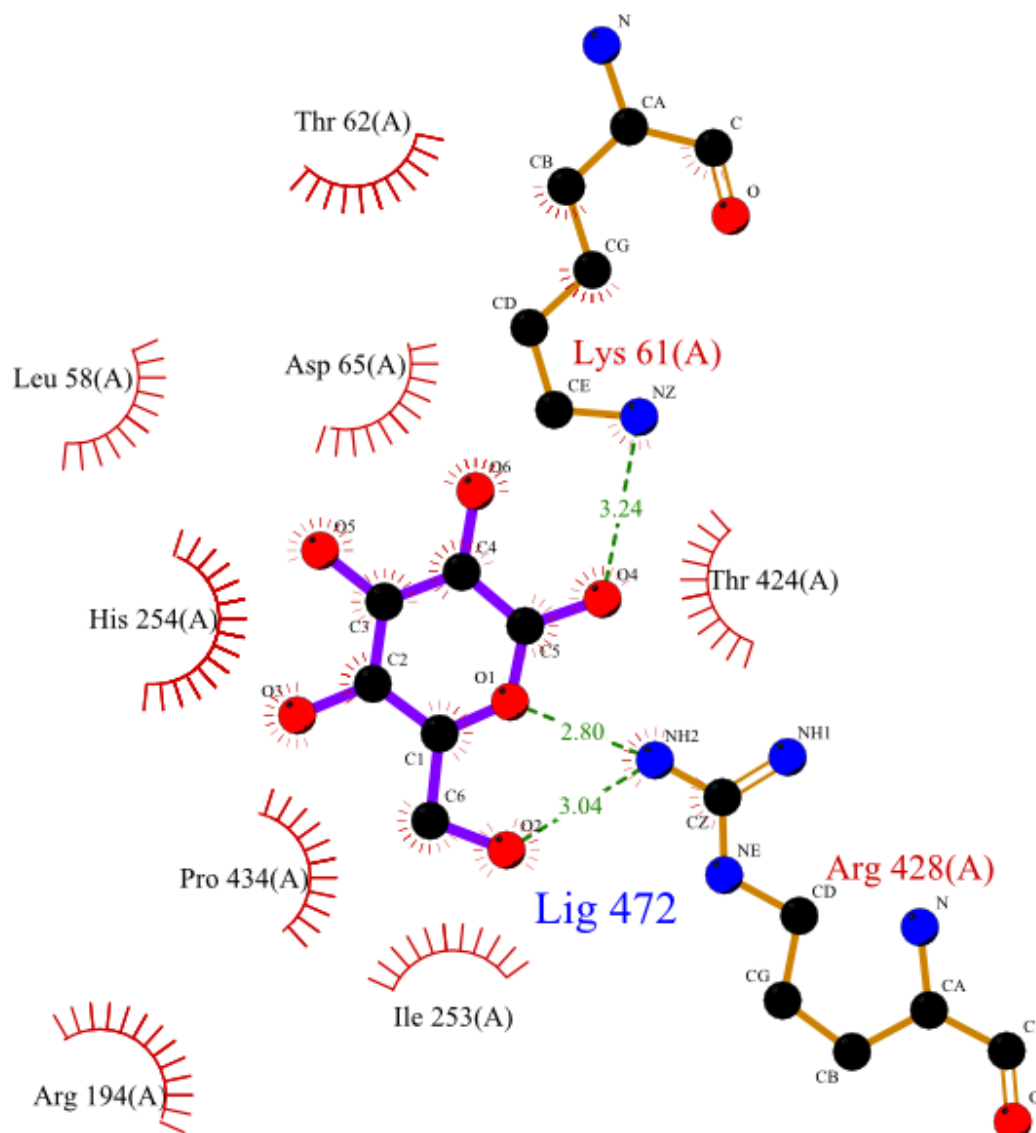


Figura 14: Formato *Postscript* do Ligplot. Exemplo de interação da proteína da β -glucosidase de código PDB:1GAM com o ligante “glicose”.

Fonte: Gerado pelo programa Ligplot.

vos com diferentes formatos de sequência e alinhamentos múltiplos de sequência. Além disso, possui módulos que lidam com estruturas macromoleculares terciárias, realizam acesso a banco de dados disponíveis na *web* e também fornecem métodos numéricos para a aprendizagem estatística.

Desde a sua criação em 1999, *Biopython* tem crescido, apresentando uma grande coleção de módulos atualmente. Essa ferramenta destina-se a programadores da biologia computacional que necessitam incorporar em seus *scripts* ou em seu próprio *software* módulos que ajudem em seus problemas específicos. A figura 15 apresenta um exemplo do uso da ferramenta *Biopython* discutida nessa seção para a transformação de uma cadeia de DNA em uma cadeia de RNA através do módulo *Bio.Seq* - função *gene.transcribe()* da biblioteca *Biopython*.

```
>>> from Bio.Seq import Seq
>>> from Bio.Alphabet import generic_dna
>>> gene = Seq("ATGAAAGCAATTTTCGTACTG"
...           "AAAGGTTGGTGGCGCACTTGA",
...           generic_dna)
>>> print gene.transcribe()
AUGAAAGCAAUUUUCGUACUGAAAGGUUGGUGGCGCACUUGA
```

Figura 15: Exemplo de uso da biblioteca *Biopython*

3.3.2 MySQL

MySQL² é um sistema gerenciador de banco de dados de código aberto. Utiliza a linguagem SQL (do inglês *Structured Query Language*) como interface. Além de ser um *software* livre, possui facilidade no manuseio, excelente desempenho e estabilidade, e é pouco exigente quanto a recursos de *hardwares*.

3.3.3 Pymol

Segundo (SCHRÖDINGER, 2010) *PyMOL*³ é um sistema de visualização molecular, criado inicialmente por Warren Lyford DeLano e comercializado por DeLano Scientific LLC, que era uma empresa de *software* privada dedicada à criação de ferramentas úteis que se tornaram universalmente acessíveis às comunidades científicas e educacionais. Atualmente é comercializada por Schrödin, Inc. *PyMOL* produz imagens tridimensionais de alta qualidade de pequenas moléculas e de macromoléculas maiores, como proteínas. *PyMOL* é umas das poucas ferramentas de visualização de código aberto disponíveis para uso em biologia computacional (SCHRÖDINGER, 2010).

²<http://www.oracle.com/br/products/mysql/overview/index.html>

³<https://www.pymol.org/>

3.3.4 Cytoscape

*Cytoscape*⁴ é uma plataforma de código aberto para a visualização de redes de interação molecular e vias biológicas, a qual integra essas redes com anotações, perfis de expressão gênica, além de outros dados de estado. Embora *Cytoscape* tenha sido projetado para a pesquisa biológica, atualmente é uma plataforma geral para análise complexa de rede e visualização. A distribuição do *Cytoscape* oferece um conjunto básico de funcionalidades para integração de dados, análise e visualização. Funcionalidades adicionais estão disponíveis como *plugins*. Esses *plugins* são utilizados para as redes e análise de perfil molecular, novos *layouts*, suporte adicional para formato de arquivo, além de *scripts* e conexão com banco de dados. *Plugins* para o *Cytoscape* podem ser desenvolvidos por qualquer pessoa utilizando a API de código aberto do *Cytoscape*, baseada em Java. O desenvolvimento de *plugins* é incentivado na comunidade do *Cytoscape*.

⁴http://www.cytoscape.org/what_is_cytoscape.html

4 TRABALHOS RELACIONADOS

Neste capítulo, são apresentadas as ferramentas relacionadas com a desenvolvida neste trabalho. Tem como objetivo discutir e comparar as funcionalidades implementadas com as existentes em outras ferramentas. No final deste capítulo, é apresentada uma tabela que compara as funcionalidades da ferramenta implementada *Cat-p-Data* neste trabalho com as demais estudadas.

4.1 Banco de dados FReDD (Supporting Mining Strategies through a Flexible-Receptor Docking Database)

Nesta seção é apresentado o banco de dados *FReDD* (do inglês, *Flexible-Receptor Docking Database*), o qual é um banco de dados abrangente e robusto que é capaz de armazenar resultados de docagem molecular com um certo nível de detalhamento e comporta uma quantidade de dados significativa. Segundo (WINCK et al., 2009) o banco de dados *FReDD* foi desenvolvido com o objetivo de armazenar resultados completos de docagem molecular mostrando relações entre suas características. Os dados armazenados neste banco podem ser adquiridos de diversas fontes como por exemplo, um experimento de docagem molecular com receptor flexível o qual envolve características de átomos e resíduos de receptores e ligantes.

Outra característica importante do banco de dados *FReDD* é disponibilizar funcionalidades para fornecer uma rápida e eficiente geração de entrada para a mineração de dados. Esses dados de entrada para a mineração são gerados no formato do software *Weka*. O banco de dados *FReDD* é composto por 16 tabelas conforme é apresentado na Figura 16 (WINCK et al., 2010).

É importante ressaltar que manipulações de dados através das ações de inserção, seleção e exclusão são produzidas através de *scripts SQL* (*Structured Query Language*) implementados manualmente pelos próprios usuários do banco de dados, o que implica em um conhecimento prévio de programação da parte do pesquisador.

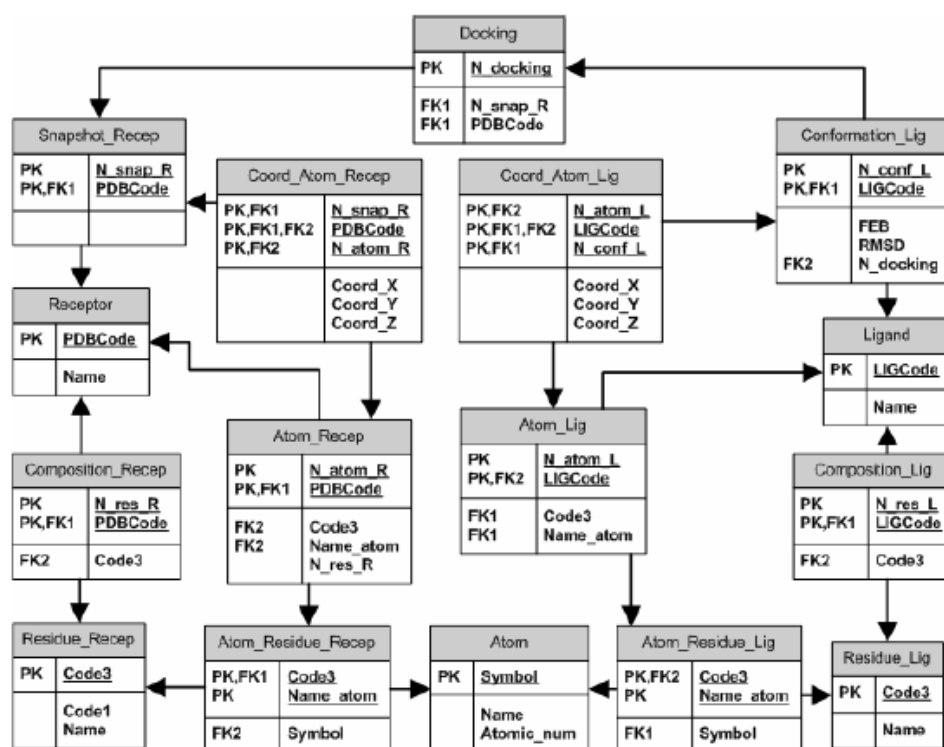


Figura 16: Esquema do banco de dados FReDD

Fonte: artigo intitulado “FReDD: Supporting Mining Strategies through a Flexible-Receptor Docking Database”

4.2 The PDBbind DataBase

O banco de dados PDBbind foi desenvolvido com o objetivo de fornecer um conjunto abrangente de afinidades de ligações para os complexos proteína-ligante do *PDB* (*Protein Data Bank*). Segundo (WANG et al., 2005), o foco deste banco de dados é armazenar somente complexos do tipo proteína-ligante, complexos do tipo proteína-proteína e proteína-ácido nucleico não são considerados. Para a aquisição e classificação dos dados extraídos do *PDB* este software utiliza um módulo na linguagem de programação C++. O algoritmo de classificação pode ser visualizado através do fluxograma da Figura 17.

Outra importante característica do software é a utilização de uma interface *web* que possui integração com a base de dados *PBbind* e que é também utilizado pelos desenvolvedores do projeto para apresentar seus principais resultados. O *front-end* desta aplicação foi construído utilizando a linguagem de programação *Java*. Todos os outros programas foram integrados e alimentados utilizando *Apache* em uma plataforma *Linux Red Hat*. O banco de dados *PDBbind* foi implementado utilizando banco de dados relacional *Oracle*. O banco de dados *PDBbind* está disponível com uma licença de demonstração para o usuário no seguinte endereço: ¹.

¹ <http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp>

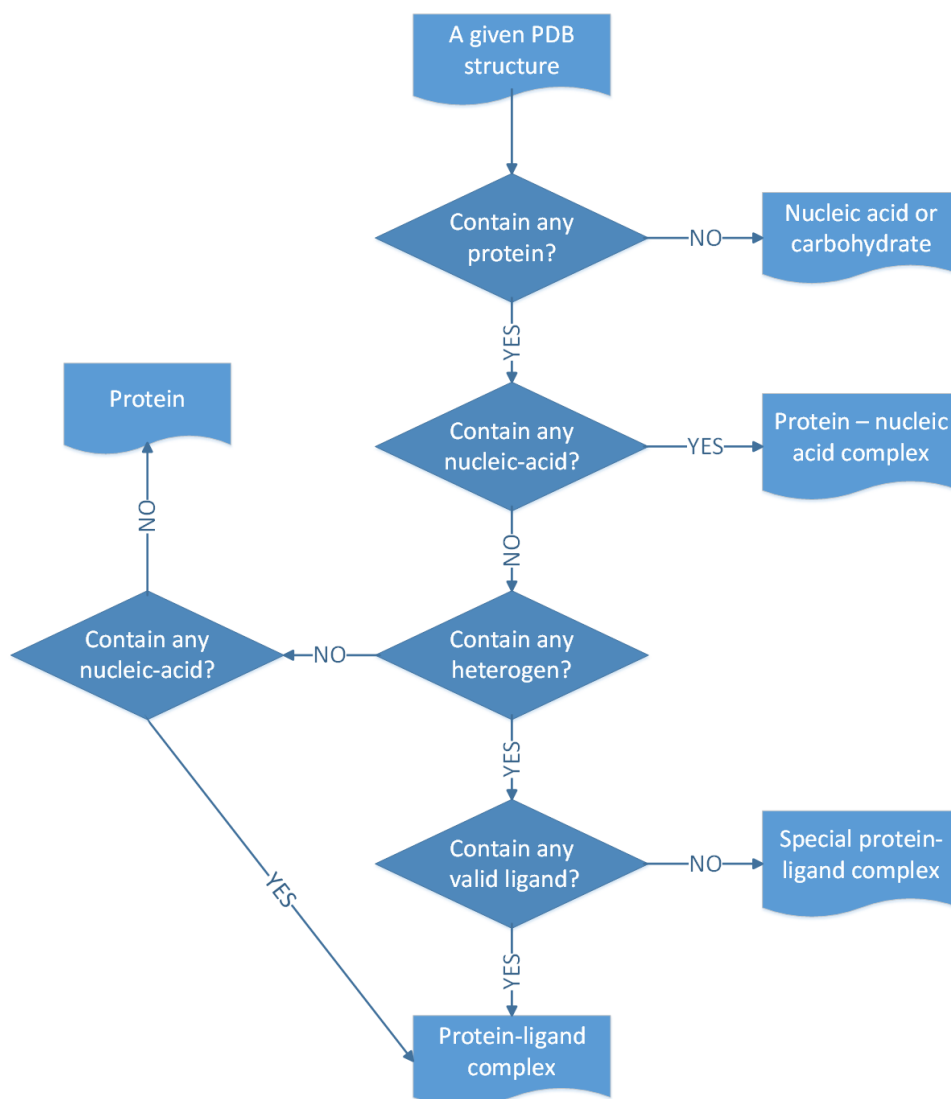


Figura 17: Esquema utilizado para a identificação e classificação dos complexos proteína-ligante.

Fonte: adaptado pelo autor do artigo intitulado “The PDBbind Database: Methodologies and Updates”

4.3 The RS-PDB Database

O trabalho descrito nesta seção consiste em um método de análise automática de estruturas depositadas no *PDB*. Segundo (SZABADKA; GROLMUSZ, 2006), o método é capaz de detectar átomos faltantes, desvios de comprimento de ligação e também identificar corretamente complexos do tipo proteína-ligante. Os resultados são armazenados em um banco de dados chamado *RS-PDB* (*Rich Structure PDB*) o qual pode-se selecionar facilmente dados de *PDB* que satisfaçam diversos conjuntos de requisitos.

O RS-PDB possui um algoritmo de pré-processamento que trabalha no formato *mmCIF* (*macromolecular Crystallographic Information File*) do *PDB* e utiliza o identificador químico InChI da IUPAC. Esse algoritmo verifica erros e inconsistências nos dados

de entrada, indica quais são os átomos faltantes, decompõem as estruturas em proteínas, ácidos-nucleicos e cadeias de polissacarídeos, assim como vários tipos de ligantes como por exemplo peptídeos, metais, etc. Além disso, realiza a distinção entre ligações do tipo covalente e não-covalente e também identifica diferentes locais de ligação.

O *RS-PDB* é dividido em três partes, e estas são apresentadas de forma simplificada a seguir:

- Parte 1 - Banco de dados de pequenas moléculas - Nesta parte do banco, pequenas moléculas são armazenadas. Para cada átomo, somente o símbolo, a carga e o modelo de coordenadas são armazenados;
- Parte 2 - Dicionário de componentes químicos do *PDB* - A parte secundária do banco de dados contém o dicionário de componentes químicos do *PDB*. Um arquivo chamado de “components.cif” pode ser adquirido do web site *RCSB*. Esse arquivo consiste nas tabelas de ligação das moléculas e grupos funcionais associados com o código do aminoácido de três letras encontrado no *PDB*;
- Parte 3 - É a principal parte do banco de dados, onde são armazenados os resultados do processamento dos dados de entrada do *PDB*. Depois de um dado de entrada ser lido corretamente pelo programa, esse dado consistirá de um polímero e moléculas de ligante e cada um terá um identificador único. Esse identificador é denotado por “mol#” nas tabelas do banco de dados.

4.4 **PDBest: a user-friendly platform for manipulating and enhancing protein structures**

Segundo (GONÇALVES et al., 2015), *PDBest* é uma plataforma amigável, disponibilizada livremente para a aquisição, manipulação e normalização de estruturas de proteínas. Possui uma interface gráfica intuitiva, que permite aos usuários sem conhecimento prévio em programação, adquirir e manipular suas estruturas proteicas. A plataforma exporta também protocolos, permitindo assim, que os usuários compartilhem facilmente buscas por *PDBs* e critérios de filtragem, melhorando a análise da reprodutibilidade.

A plataforma *PDBest* foi desenvolvida visando alcançar os seguintes objetivos: fornecer meios para melhorar a qualidade dos dados de estrutura molecular utilizados nas áreas de Biologia e Bioinformática, torná-lo acessível a um público mais amplo, fornecer escalabilidade e processamento paralelo, assim como, reprodutibilidade e reutilização. A Figura 18 apresenta a arquitetura e funcionamento do *PDBest*. Com relação ao seu funcionamento, o *PDBest* permite ao usuário selecionar proteínas através de uma consulta à base de dados do *PDB* ou através do carregamento local do arquivo de proteína.

A partir disso, o usuário pode utilizar filtros nas proteínas para adicionar ou remover hidrogênios, remover solventes, filtrar cadeias ou modelos e selecionar ligantes. Paralelamente a esse processo de filtragem, o PDBest processa e trata ocupâncias múltiplas de átomos e também átomos ou resíduos faltantes nas proteínas. Depois disso, o usuário pode exportar o protocolo de filtragem utilizado em determinado grupo de proteínas. Todas as informações processadas ficam armazenadas em um conjunto de dados disponível na ferramenta.

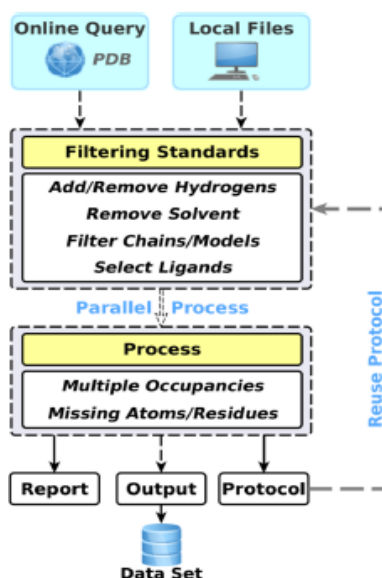


Figura 18: Arquitetura da ferramenta PDBest.

Fonte: artigo intitulado “PDBest: a user-friendly platform for manipulating and enhancing protein structures”

Segundo (GONÇALVES et al., 2015), através da arquitetura apresentada na figura 18, o *PDBest* visa oferecer uma ferramenta poderosa para os pesquisadores da área de Biologia Estrutural. Através de uma interface gráfica fornece uma eficiente maneira de adquirir, manipular e verificar dados do *PDB* como meio de melhorar a qualidade e integridade dos dados. *PDBest* está disponível em todos os principais sistemas operacionais e provou ser uma ferramenta poderosa para o estudo de estruturas proteicas.

4.5 Pre-calculated protein structure alignments at the RCSB PDB website

Nesta seção é apresentada uma ferramenta para auxiliar na comparação de alinhamentos de estruturas primárias e terciárias. Segundo (PRLIĆ et al., 2010), esta ferramenta suporta vários aplicativos no site do *RCSB PDB* e seu funcionamento é descrito da seguinte forma: primeiramente, um *web-service* de alinhamento estrutural calcula alinhamentos em pares. Em segunda instância, um aplicativo executa localmente os alinhamentos e fornece a visualização dos resultados. O terceiro passo visa calcular comparações

entre estruturas terciárias para todo o *PDB* e assim fornecer esses resultados semanalmente. A cada semana todas as proteínas são alinhadas e comparadas todas contra todas pelos servidores do *RCSB PDB* ao mesmo tempo que o banco de dados *PDB* é atualizado. Atualmente, essa atualização semanal requer um cálculo de cerca de 1 milhão de alinhamentos. Os resultados estão disponíveis no servidor *FTP* do *RCSB PDB* através do endereço “ftp://resources.rcsb.org/fatcat_rigid_pdb_all/fatcat_rigid_pdb_all.txt.gz”. A ferramenta permite também o download dos resultados no formato *XML*. Todas as informações sobre essa ferramenta estão disponíveis no endereço ².

4.6 **ProtDCal: A program to compute general-purpose numerical descriptors for sequences and 3D-structures of proteins**

Segundo (RUIZ-BLANCO et al., 2015), o crescimento exponencial das bases de dados de proteínas e de sequência estão permitindo abordagens multifacetadas para a compreensão da relação estrutura-função. Atualmente, os avanços na computação tornam possível a aplicação de técnicas de mineração de dados e reconhecimento de padrões bem estabelecidos para esses dados, com isso é possível aprender sobre modelos que efetivamente se relacionam a estrutura e função. No entanto, a extração de descritores numéricos significativos a partir de sequência e estrutura tridimensional da proteína é uma questão fundamental que requer uma solução eficiente e amplamente disponível. Dessa forma, essa seção apresenta a ferramenta *ProDCal*, um conjunto de programas capaz de gerar dezenas de milhares de características, considerando descritores baseados em estrutura secundária como também em estrutura terciária.

É demonstrado, por meio da análise do componente principal, como descritores baseados na sequência pelo *ProDCal* fornecem novas e mais informações relevantes que não são codificadas atualmente por servidores que geram características baseadas em estrutura secundária de proteínas.

Portanto, *ProDCal* fornece uma interface gráfica amigável multi-plataforma a qual foi desenvolvido na linguagem de programação Java. *ProDCal* introduz codificação local e baseada em grupo, as quais aumentam a diversidade das informações capturadas.

Este *software* destina-se a fornecer uma ferramenta útil para a codificação geral de sequências de proteínas. Dentre as várias aplicações que a ferramenta *ProDCal* possui, destacam-se classificação de proteínas, análise de similaridade e funções de predição.

4.7 **Tabela comparativa**

Nesta seção foram apresentados os trabalhos relacionados, com o objetivo de verificar quais são as características principais de cada ferramenta e dessa forma comparar em

²http://www.rcsb.org/pdb/static.do?p=general_information/cluster/structureAll.jsp

relação ao *software* desenvolvido nesse trabalho. A tabela 2, apresenta uma comparação geral entre todas as ferramentas relacionadas ao trabalho desenvolvido nessa dissertação.

Tabela 2: Comparação entre os trabalhos relacionados e a ferramenta desenvolvida nesse trabalho.

-	FReDD	PDBbind	RS-PDB	PDBest	Pre-calc protein	ProtDCal	Cat-p-data
(BD ou Ferramenta ou Ambos)	BD	BD	Ambos	Ferramenta	Ferramenta	Ferramenta	Ambos
Interface simples	Não	Sim	Não	Sim	Sim	Sim	Sim
Tipos de dados (Sequência)	-	-	-	-		V	V
Tipos de dados (Estrutura)	V	V	V	V	V	V	V
Tipos de dados (Docking e triagem virtual)	V	-	-	-	-	-	V
Tipos de dados (ligantes)	-	V	V	-	-	-	V

A ferramenta *Cat-p-data* tem por objetivo instanciar um conjunto específico de moléculas do PDB, através da aquisição dessas moléculas pela própria ferramenta, a qual possui conexão com *web service RESTful*. Assim, esses dados proteicos são armazenados no banco de dados local da ferramenta e vinculados a um projeto único. Com relação à análise de dados, o *Cat-p-data* possui funcionalidades de alinhamento de sequência, módulo de estudo de similaridade, alinhamento estrutural, ambiente de visualização de proteínas, módulo de visualização dos ligantes e módulos para a atualização do conjunto de proteínas. Por exemplo, com o uso dos módulos de alinhamento sequencial e módulo de similaridade, pode-se destacar a possibilidade de filtrar um conjunto mais específico de proteínas, as quais compartilhem uma determinada similaridade em sua estrutura primária e, dessa forma, eliminar as que não possuem. Esses dois módulos utilizados em conjunto, possibilitam ao pesquisador ter um conjunto de dados mais específico e autêntico para a realização de sua pesquisa.

5 RESULTADO 1: FERRAMENTA CAT-P-DATA (CUSTOM ANALYSIS TOOL FOR PROTEIN DATA)

Bancos de dados públicos como o *PDB* (*Protein Data Bank*) (BERMAN et al., 2000), *NCBI protein*¹ e *ZINC* (IRWIN et al., 2012) possuem uma enorme quantidade de dados armazenados. Com esta grande quantidade de dados, fica evidente que o uso de ferramentas que realizam a aquisição e análise desse conjunto de dados é uma necessidade. Além disso, relacionar dados estruturais de proteínas com dados sequenciais de proteínas e relacionar experimentos de docagem molecular e triagem virtual ainda é um desafio. Dessa forma, é apresentado nesse capítulo a ferramenta proposta nesse trabalho *Cat-p-Data* (*do inglês Custom Analysis Tool for Protein Data*). Esse capítulo apresenta detalhadamente todos os módulos, funcionalidades e arquitetura da ferramenta proposta nesse trabalho.

5.1 Arquitetura e módulos do Cat-p-data

5.1.1 Módulos do Cat-p-data

Nesta sub-seção é apresentado os módulos da aplicação desenvolvida neste trabalho. A Figura 19 apresenta cada módulo do sistema. Com relação ao diagrama da Figura 19, cada módulo da aplicação é discutido a seguir:

- Cat-p-data - Esse módulo representa a aplicação em si, onde o banco de dados local e a interface estão localizados nesse esquema;
- PDB - Esse módulo representa o banco de dados *web* público *PDB* (*Protein Data Bank*). A aplicação conecta-se com esse banco para a aquisição de estruturas moleculares através do *web-service RestFull* fornecido pelo próprio *PDB*. A partir dessa aquisição dos dados, o usuário cria um projeto único onde todos esses dados ficam vinculados;
- Ferramenta de alinhamento de sequências - Esse módulo representa a ferramenta para alinhamento de sequências. Com essa funcionalidade é permitido ao usuário

¹<http://www.ncbi.nlm.nih.gov/protein>

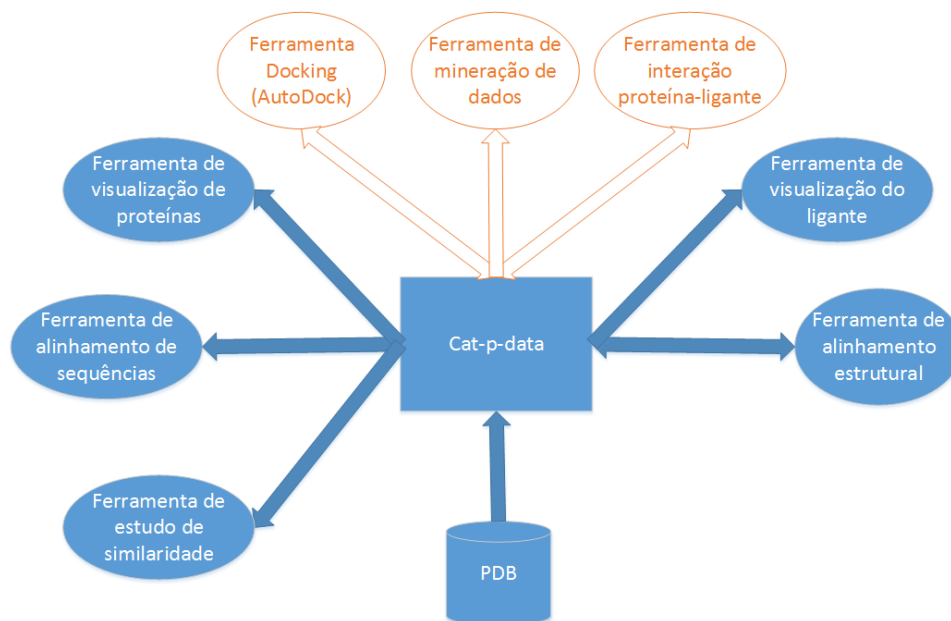


Figura 19: Módulos do sistema

alinhar as sequências de aminoácidos (estrutura primária) todas com todas proteínas do seu projeto e assim visualizar aquelas que possuem similaridade ou não. Essa funcionalidade deve ser executada primeiramente, pois a maioria das outras funcionalidades dependem desta para ter êxito;

- Ferramenta de visualização de proteínas - Esse módulo representa a ligação do *Cat-p-data* com a ferramenta de visualização de proteínas. Com essa ferramenta é possível visualizar as proteínas em um ambiente tridimensional. Para isso, foi utilizado a biblioteca gráfica *Pymol*.
- Ferramenta de visualização dos ligantes - Esse módulo lista todos os ligantes presentes nas estruturas das proteínas vinculadas a um projeto;
- Ferramenta de alinhamento estrutural - Esse módulo representa a ferramenta para alinhamento da estrutura terciária e quaternária. Com relação ao seu funcionamento, primeiramente o usuário seleciona uma das proteínas de um dos seus projetos para servir de base. A partir deste ponto, todas as proteínas do projeto carregado são alinhadas com relação à estrutura da proteína base anteriormente selecionada. Essa funcionalidade é importante, por exemplo, para preparar moléculas para experimentos de *Virtual Screening (VS)*. Com o alinhamento estrutural todas as proteínas serão alinhadas em um mesmo plano de coordenadas cartesianas, isso significa que, o pesquisador terá que selecionar uma única região no plano tridimensional para o ligante atracar nas diferentes moléculas;
- Ferramenta de estudo de similaridade. Permite ao usuário verificar a relação de

similaridade na estrutura sequencial entre proteínas de um mesmo projeto. Essa ferramenta gera um arquivo no formato de entrada para o *software Cytoscape*, o qual gera grafos de fácil interpretação.

Os três módulos a seguir estão em desenvolvimento em outro trabalho. Quando finalizados, serão inseridos na ferramenta desenvolvida nesse trabalho.

- Ferramenta docking (AutoDock) - Esse módulo representa a ferramenta que armazena resultados de docagem molecular no banco de dados local;
- Ferramenta de mineração de dados - Prepara entradas para mineração de dados;
- Ferramenta interação proteína-ligante (ligplot) - Esse módulo representa a ferramenta que possibilita ao usuário visualizar os resíduos ou átomos de proteínas e ligantes que fizeram contato em um determinado experimento de docagem molecular.

5.1.2 Arquitetura do Cat-p-Data

Nesta sub-seção é apresentada a arquitetura da aplicação *Cat-p-data*. A Figura 20 a seguir exemplifica como é dado o funcionamento de todo o sistema.

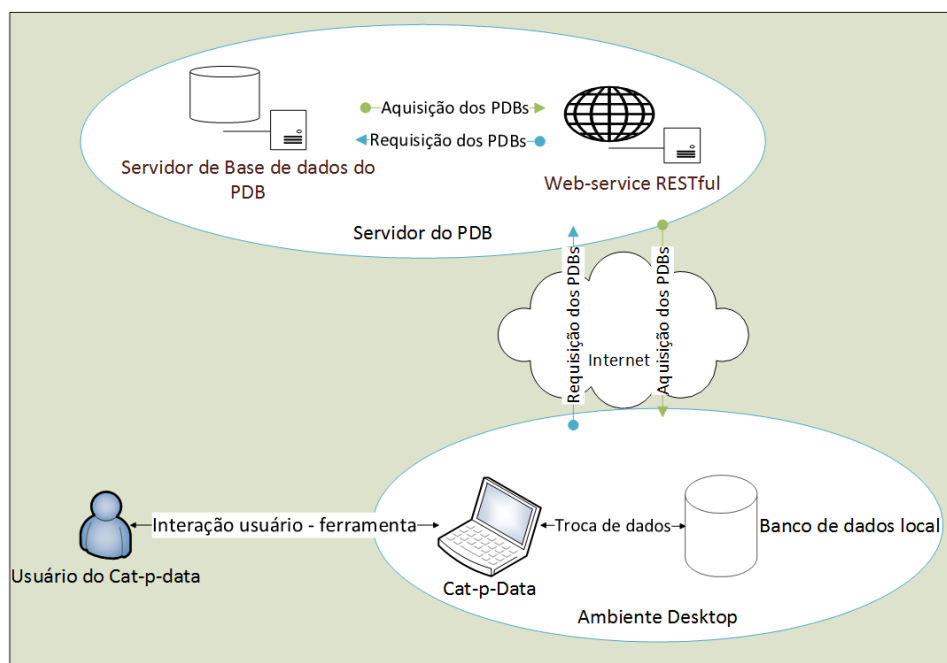


Figura 20: Arquitetura do Cat-p-data

Como pode ser visto na Figura 20, a ferramenta desenvolvida neste trabalho executa em um ambiente desktop. A partir da interface do programa, o usuário busca o seu conjunto específico de proteínas através de uma palavra chave em um campo de busca da ferramenta. A seguir, um conjunto de possíveis proteínas é apresentado na ferramenta por

meio de uma lista de *PDB IDs* e assim o usuário pode realizar a aquisição destas proteínas. De acordo com a Figura 20, pode-se notar que esse processo de aquisição das proteínas é dado através do uso da *internet* em conjunto com o *web-service Restfull* provido pelo próprio PDB. Com esse *web-service* disponível na *web*, a ferramenta *Cat-p-Data* conecta-se ao mesmo e realiza a aquisição dos dados proteicos. A partir deste ponto, sem mais o uso de conexão com a internet, a ferramenta realiza a inserção destes dados no banco de dados local discutido na seção 5.2 e apresentado na figura 20 dentro do ambiente *desktop*. A seguir, com a inserção das estruturas moleculares no banco de dados local finalizada, o usuário pode utilizar as diversas funcionalidades que a ferramenta *Cat-p-Data* oferece.

5.2 Banco de dados local

Nesta seção é apresentado o banco de dados local que faz parte da ferramenta proposta neste trabalho. O objetivo desse banco de dados é proporcionar o acesso as características das estruturas moleculares em um ambiente *off-line*. Esse banco de dados é populado após toda vez que o usuário realizar a aquisição de um conjunto de dados específico a partir da ferramenta *Cat-p-data*. A Figura 21 apresenta o modelo entidade relacionamento desse banco de dados.

A base de dados *AAindex* discutida na seção 3.1.4, foi convertida para uma tabela de um modelo relacional de banco de dados, para facilitar a busca por informações em trabalhos futuros. As figuras 22 e 23 apresentam o formato da base de dados após a conversão. Como pode ser visto na Figura 22, a tabela “properties” do modelo relacional do banco *AAIndex* contém os campos “id_property”, “code” e “description”, além dos outros 20 campos que correspondem aos 20 aminoácidos que recebem um valor numérico para cada propriedade e que são apresentados na Figura 23. O campo “id_property” possui um identificador numérico único para a propriedade, o campo “code” possui um identificador único do tipo “varchar” que representa um código para determinada propriedade e o campo “description” do tipo “varchar” possui uma descrição para determinada propriedade. Essa tabela (representada pelas figuras 22 e 23) pode ser interpretada da seguinte maneira, as linhas representam propriedades dos aminoácidos (um total de 544) e as colunas representam os valores de cada propriedade para cada um dos 20 tipos de aminoácidos. Parar exemplificar a relação entre colunas e linhas da tabela da figura 22, destaca-se que os 20 valores da linha 1 são os valores que cada aminoácido retém em relação a propriedade de código(*code*) igual a “ANDN920101” da tabela da figura 22 e assim sucessivamente.

A especificação sobre cada tabela do banco de dados da figura 21 é apresentada a seguir.

- **Residue_aaindex** - Tabela fixa que armazena informações sobre as diversas propriedades dos aminoácidos. Esta tabela é fixa e armazena 20 aminoácidos onde cada um

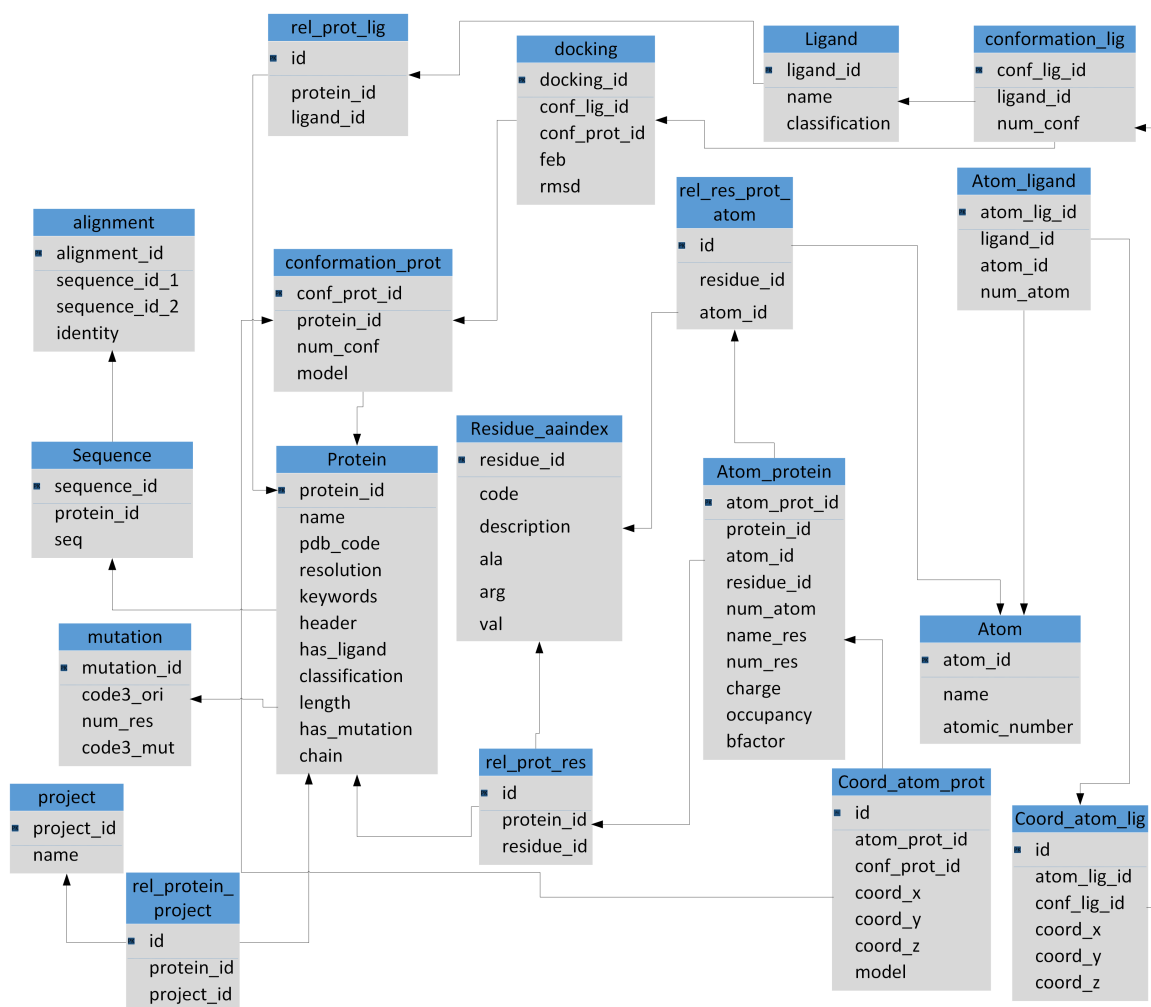


Figura 21: Modelagem ER do banco de dados local para o armazenamento dos dados das proteínas.

id_property	code	description
5	ANDN920101	alpha-CH chemical shifts (Andersen et al., 1992)
6	ARGP820101	Hydrophobicity index (Argos et al., 1982)
7	ARGP820102	Signal sequence helical potential (Argos et al., 1982)
8	ARGP820103	Membrane-buried preference parameters (Argos et al., 1982)
9	BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx, 1975)
10	BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx, 1975)
11	BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx, 1975)
12	BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
13	BIGC670101	Residue volume (Bigelow, 1967)
14	BIOV880101	Information value for accessibility; average fraction 35% (Biou et al., 1988)
15	BIOV880102	Information value for accessibility; average fraction 23% (Biou et al., 1988)
16	BROC820101	Retention coefficient in TFA (Browne et al., 1982)
17	BROC820102	Retention coefficient in HFBA (Browne et al., 1982)
18	BULH740101	Transfer free energy to surface (Bull-Breese, 1974)

Figura 22: Campos *code* e *description* da tabela *properties* do BD AAIndex após a conversão para o modelo relacional da base de dados proposta.

Elaborado pelo autor.

ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
4.35	4.38	4.75	4.76	4.65	4.37	4.29	3.97	4.63	3.95	4.17	4.36	4.52	4.66	4.44	4.50	4.35	4.7	4.6	3.95
0.61	0.60	0.06	0.46	1.07	0.	0.47	0.07	0.61	2.22	1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32
1.18	0.20	0.23	0.05	1.89	0.72	0.11	0.49	0.31	1.45	3.23	0.06	2.67	1.96	0.76	0.97	0.84	0.77	0.39	1.08
1.56	0.45	0.27	0.14	1.23	0.51	0.23	0.62	0.29	1.67	2.93	0.15	2.96	2.03	0.76	0.81	0.91	1.08	0.68	1.14
1.	0.52	0.35	0.44	0.06	0.44	0.73	0.35	0.60	0.73	1.	0.60	1.	0.60	0.06	0.35	0.44	0.73	0.44	0.82
0.77	0.72	0.55	0.65	0.65	0.72	0.55	0.65	0.83	0.98	0.83	0.55	0.98	0.98	0.55	0.55	0.83	0.77	0.83	0.98
0.37	0.84	0.97	0.97	0.84	0.64	0.53	0.97	0.75	0.37	0.53	0.75	0.64	0.53	0.97	0.84	0.75	0.97	0.84	0.37
0.357	0.529	0.463	0.511	0.346	0.493	0.497	0.544	0.323	0.462	0.365	0.466	0.295	0.314	0.509	0.507	0.444	0.305	0.42	0.386
52.6	109.1	75.7	68.4	68.3	89.7	84.7	36.3	91.9	102.0	102.0	105.1	97.7	113.9	73.6	54.9	71.2	135.4	116.2	85.1
16.	-70.	-74.	-78.	168.	-73.	-106.	-13.	50.	151.	145.	-141.	124.	189.	-20.	-70.	-38.	145	53	123
44.	-68.	-72.	-91.	90.	-117.	-139.	-8.	47.	100.	108.	-188.	121.	148.	-36.	-60.	-54.	163	22	117
7.3	-3.6	-5.7	-2.9	-9.2	-0.3	-7.1	-1.2	-2.1	6.6	20.0	-3.7	5.6	19.2	5.1	-4.1	0.8	16.3	5.9	3.5
3.9	3.2	-2.8	-2.8	-14.3	1.8	-7.5	-2.3	2.0	11.0	15.0	-2.5	4.1	14.7	5.6	-3.5	1.1	17.8	3.8	2.1
-0.20	-0.12	0.08	-0.20	-0.45	0.16	-0.30	0.00	-0.12	-2.26	-2.46	-0.35	-1.47	-2.33	-0.98	-0.39	-0.52	-2.01	-2.24	-1.56

Figura 23: Continuação da tabela *properties* do banco AAindex.
Elaborado pelo autor.

possui 544 propriedades. É uma variação do base de dados AAindex apresentada na seção 3.1.4. Seu conteúdo está em parte descrito nas Figuras 22 e 23;

- Atom - armazena todos os átomos existentes na literatura. É uma tabela fixa;
- project - armazena o nome do projeto. A cada conjunto de dados proteicos adquirido pelo usuário é associado a um determinado nome para o projeto. O usuário escolhe este nome;
- Protein - armazena dados como o nome da proteína, código PDB, cabeçalho do PDB, palavras-chave, nome do ligante e a existência dele ou não na proteína (campo *has_ligand*), todas cadeias existentes no arquivo PDB, entre outros campos;
- mutation - armazena informações sobre mutações das moléculas.
- rel_protein_project - armazena o relacionamento de proteínas com um determinado projeto. Uma proteína pode estar vinculada a diferentes projetos. O código identificador da proteína (*protein_id*) é vinculado a vários projetos na tabela;
- conformation_prot - armazena diferentes conformações que uma proteína possui através de simulações de dinâmica molecular, além disso, armazena conformações de uma proteína do *PDB* que foram depositadas utilizando as técnicas de Ressonância Magnética Nuclear (do inglês Nuclear Magnetic Resonance) e Cristalografia por Raio X (do inglês X-Ray Crystallography);
- rel_prot_res - armazena o relacionamento entre as tabelas “Protein” e “Residue_aaindex”. Através desta tabela (*rel_prot_res*) é possível identificar todos os aminoácidos presentes em determinada proteína;
- rel_res_prot_atom - armazena o relacionamento entre as tabelas “Residue_aaindex” e “Atom_protein”. Através desta tabela pode-se analisar todos os átomos dos resíduos presentes em determinada proteína;

- Atom_protein - armazena dados referentes aos átomos pertencentes a determinada proteína;
- Coord_atom_prot - armazena as coordenadas dos átomos nas proteínas;
- Sequence - armazena a estrutura primária das proteínas de um projeto.
- alignment - armazena resultados de alinhamento sequencial entre todas as proteínas de um ou mais projetos.
- Ligand - armazena dados como nome do ligante e classificação. Esta tabela armazena os nomes dos ligantes presentes nas proteínas vinculadas aos projetos;
- Atom_ligand - armazena dados referentes aos átomos pertencentes a determinado ligante;
- Coord_atom_lig - ; armazena as coordenadas dos átomos no ligante;
- conformation_lig - armazena diferentes conformações que um ligante pode possuir através de processos de docagem molecular;
- rel_prot_lig - armazena o relacionamento entre proteínas e ligantes;
- docking - armazena informações sobre resultados de docagem molecular. Não está sendo utilizada neste trabalho. Com a inserção do módulo de docagem molecular (trabalho futuro) esta será utilizada;

5.3 Funcionalidades do Cat-p-data.

Com relação à ferramenta, seu funcionamento pode ser exemplificado através de duas etapas principais. A primeira etapa de funcionamento da ferramenta, possibilita o usuário a adquirir as estruturas moleculares através de uma conexão com a internet em conjunto com a interface da mesma. O usuário da ferramenta pode realizar a aquisição de diversos conjuntos diferentes de proteínas e salvá-los como projetos. A partir deste ponto, sem mais o uso de conexão com a internet, a segunda etapa de funcionamento da ferramenta implementa algumas funcionalidades que possibilitam o pesquisador a investigar características com relação a estrutural sequencial (1D) e estrutura terciária (3D) das proteínas. Além dessas funcionalidades, o usuário pode utilizar um ambiente gráfico tridimensional para a visualização das moléculas de determinado projeto.

Acredita-se que com esse ferramental, o pesquisador poderá realizar uma série de manipulações nas moléculas de acordo com o objetivo de sua pesquisa, trabalhando em um ambiente exclusivo e com dados específicos para a realização de seu trabalho. Todas as funcionalidades já disponíveis são exemplificadas nas seções à seguir através de resultados já gerados pela ferramenta. Para a validação das funcionalidades da ferramenta,

será utilizado o conjunto de estruturas macromoleculares da enzima β -glicosidase. Todas essas estruturas de proteínas foram adquiridas por meio da interface do *software* e armazenadas no banco de dados local do mesmo.

As funcionalidades implementadas na ferramenta *Cat-p-Data* estão resumidas na Figura 19 e descritas a seguir:

- Aquisição dos dados e criação de projeto. Aquisição e armazenamento dos dados de proteína no banco de dados local, possibilitando o usuário a trabalhar em um ambiente *off-line* após a aquisição dos conjuntos de moléculas;
- Alinhamento sequencial. É permitido ao usuário realizar alinhamento sequencial entre todas as moléculas do projeto;
- Ferramenta de estudo de similaridade. Permite ao usuário verificar a relação de similaridade na estrutura sequencial entre proteínas de um mesmo projeto. Essa ferramenta gera um arquivo no formato de entrada para o *software Cytoscape*, o qual gera grafos de fácil interpretação;
- Alinhamento estrutural. É permitido ao usuário alinhar várias proteínas do projeto em relação a sua estrutura tridimensional. O usuário considera como base uma das proteínas de um projeto e todas as outras selecionadas serão alinhadas com a base. Esse módulo utiliza o resultado do alinhamento sequencial e utiliza heurísticas criadas para otimizar o resultado do alinhamento.
- Ambiente tridimensional de visualização de proteínas. A ferramenta utiliza a biblioteca gráfica *Pymol* para fornecer ao usuário a possibilidade de visualização em um ambiente tridimensional;
- Ferramenta de visualização do ligante. Possibilita ao usuário identificar quais são os ligantes pertencentes às proteínas de um projeto. Este módulo lista todas as ligantes de todas as proteínas de um projeto, apresentando a quantidade de ligantes no mesmo. Além disso, é permitido ao usuário selecionar uma ou mais proteínas e assim listar somente os ligantes pertencentes a estas;
- Módulo de atualização do conjunto de proteínas. Utilizado para buscar novas proteínas no *PDB* utilizando a mesma palavra-chave usada na criação do projeto;
- Módulo de limpeza do conjunto de proteínas. Utilizado para eliminar possíveis proteínas que não pertençam ao real grupo desejado. Devido ao fato do uso de uma palavra-chave no campo de busca da ferramenta, algumas proteínas que retornam não são proteínas que o usuário deseja ter no seu projeto, assim este módulo serve para eliminá-las;

- Módulo de edição do conjunto de proteínas. Utilizado para vincular proteínas de um projeto para outro existente. Se uma proteína já está armazenada em um determinado projeto, o usuário poderá vincular esta a um outro projeto existente.
- Módulo de inserção de novas proteínas. Um pouco diferente do módulo de atualização que adquire todas as novas proteínas existentes no *PDB* referente à palavra-chave utilizada na criação do projeto, este módulo permite ao usuário vincular ao seu projeto qualquer proteína encontrada no *PDB* utilizando qualquer palavra-chave.

A seguir, cada funcionalidade do *Cat-p-data* é apresentada detalhadamente.

5.3.1 Funcionalidade 1: Aquisição dos dados e criação do projeto

Um dos objetivos da ferramenta *Cat-p-data* é instanciar grupos de moléculas do *PDB* para o banco de dados local da mesma. Por meio de uma palavra-chave informada pelo usuário no campo de busca da ferramenta, moléculas são listadas para aquisição. Através da interface, o usuário pode selecionar quais moléculas serão adquiridas e, depois disso, esse grupo de proteínas será vinculado a um projeto de nome único informado pelo próprio usuário. Todos esses dados extraídos dos arquivos de proteínas (átomos, coordenadas dos átomos, resíduos, ligantes, etc) são armazenados no banco de dados local da ferramenta, o que permite ao usuário trabalhar sem mais a necessidade de internet após a aquisição das proteínas. Importante ressaltar que os arquivos de proteínas adquiridos são armazenados no diretório do programa. Na aquisição das proteínas, além dos arquivos *PDBs*, são adquiridos também os arquivos no formato *FASTA* das proteínas, o qual possui a estrutura primária da proteína. Essa estrutura primária é salva também na base de dados local da ferramenta.

5.3.2 Funcionalidade 2: Alinhamento sequencial

O principal objetivo da funcionalidade de alinhamento sequencial implementada na ferramenta deste trabalho é realizar o alinhamento entre várias sequências de um mesmo projeto. Com esta funcionalidade é possível visualizar neste conjunto específico de proteínas, por exemplo, quais estruturas primárias são similares e quais não são. A partir deste conhecimento adquirido, o pesquisador poderá eliminar algumas estruturas que não lhe interessam devido a pouca similaridade com o restante do conjunto, por exemplo. O alinhamento sequencial utiliza o algoritmo *pairwise2* da biblioteca *Biopython*. O algoritmo *pairwise2* realiza o alinhamento sequencial de forma global par a par e utiliza programação dinâmica em sua implementação. Este algoritmo pode ser utilizado de duas formas, a primeira considera um *score* comum no alinhamento, ou seja, entre duas sequências de aminoácidos se aparecer aminoácido igual, adiciona um ponto ao *score*. A outra forma utiliza a matriz “blosum”, que segundo (VERLI, 2014) é gerada

The screenshot displays the CAT-P-DATA (beta-glucosidase-2015-19-05-18-52) interface. The main window is divided into several sections:

- Proteins (205):** A list of protein entries with their IDs and descriptions. The descriptions include details about the source organism (e.g., *Bacillus polymyxa*, *Trifolium repens*), the type of structure (e.g., crystal structure), and the specific protein variant (e.g., 1BGA, 1BGG, 1CBG, 1E1E, 1E1F, 1E4I, 1E4L, 1E4N, 1E5S, 1E56).
- Protein description:** A detailed view of a selected protein, showing its name, source, and a brief description of its function and structure.
- Information about ligands(2544):** A section detailing the ligands associated with the protein, including the protein name, the number of ligands, and the specific ligand names (e.g., GCO, PSG, G2F).
- Sequence Alignment:** A section showing the alignment of two protein sequences. The sequences are:


```

      2KSM.A 54 VDPVRSLSDFSSAEPVFTASV-----PIPDF
      |||. . .|| |||
      2VP8.A190 VDAVISQVT-AAERAAGVAREKVLIDPAHDF
      
```
- 3D Visualization:** A 3D ribbon diagram of a protein structure, showing the complex folding and interactions within the protein.
- 2D Visualization:** A 2D representation of the protein structure, showing the arrangement of atoms and the overall shape of the molecule.

Arrows indicate the flow of information from the protein list to the detailed views and visualizations.

Figura 24: Interface da tela de projetos da ferramenta *Cat-p-Data* apresentando algumas funcionalidades.

através de estudos de variação proteica e estão diretamente relacionadas à probabilidade de substituição de um aminoácido por outro. É importante destacar que, o alinhamento sequencial para cada projeto criado pelo usuário é executado uma única vez considerando todas as proteínas. As informações de alinhamento são automaticamente armazenadas no banco de dados local da ferramenta (Tabela *alignment* da Figura 21).

Outra possibilidade que a ferramenta de alinhamento sequencial permite, é gerar uma matriz com o mesmo formato da Figura 33, porém com valores adquiridos da matriz “blosum”. Segundo (VERLI, 2014), as matrizes “blosum” são as mais disseminadas e aplicadas para os mais diversos casos de comparação entre sequências de aminoácidos.

Além das funcionalidades já discutidas, a funcionalidade de alinhamento sequencial, ainda conta com um diferencial. A informação sobre a maior sequência de aminoácidos encontrada em um alinhamento sequencial entre duas proteínas é armazenada no banco de dados local. O algoritmo implementado sempre busca a sequência máxima entre duas estruturas primárias. A informação obtida através dessa funcionalidade é importante para o alinhamento estrutural, pois com esses dados sobre a sequência, o módulo de alinhamento estrutural utiliza dessa informação para conseguir um melhor resultado. Para essa implementação, antes do armazenamento é realizada uma seleção do aminoácido inicial e final existente nas duas proteínas alinhadas no momento. Essa seleção foi implementada através do uso de heurísticas. Essas heurísticas são apresentadas a seguir:

- Se o resultado do alinhamento sequencial entre duas proteínas em relação à similaridade de aminoácidos for menor que 50%, então é armazenado no banco de dados como aminoácido inicial e final um grupo contendo 80 aminoácidos, por exemplo, aminoácido inicial na posição 20 e final na posição 100;
- Se o resultado do alinhamento sequencial entre duas proteínas em relação à similaridade de aminoácidos for maior ou igual a 50%, então procura-se nas duas sequências de proteínas em questão o maior grupo de sequência existente considerando ambas e, assim, esses valores são armazenados no banco de dados local;
- Se o resultado do alinhamento sequencial entre duas proteínas em relação à similaridade de aminoácidos for maior ou igual a 100, então é armazenado no banco de dados como aminoácido inicial e final um intervalo contendo 80 aminoácidos. Para esse caso, o intervalo armazenado pode ser qualquer um, pois, as duas proteínas que estão sendo alinhadas compartilham 100% de similaridade em sua sequência de aminoácidos.

Para exemplificar o uso dessas heurísticas no alinhamento sequencial, a Figura 25 mostra destacado na cor cinza o maior segmento encontrado entre duas estruturas primárias. Nesse caso, as duas estruturas compartilham 92% de similaridade, e dessa forma seguindo as regras criadas, o procedimento implementado procura nas duas

```

>2H5Q:A|PDBID|CHAIN|SEQUENCE
MVSKEEDNMATIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTGGPLPFAWDILSPQFMYGSKAYVKH
PADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPDGPVMQKKTMGWEASSERMYPEDG
ALKGEIKQRLKLDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTGGMDELYK
>4ZIN:A|PDBID|CHAIN|SEQUENCE
AIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTGGPLPFAWDILSPQFMYGSKAYVKHPADIPDYLKL
SFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPDGPVMQKKTMGXEASSERMYPEDGALKGEIKQRL
KLDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTG

```

Figura 25: Alinhamento sequencial entre duas estruturas primárias. Destaque na cor cinza para o maior segmento de aminoácidos encontrado entre as duas sequencias.

sequências o maior grupo de aminoácidos existente e armazena no banco de dados local o valor inicial e final dos resíduos.

5.3.3 Funcionalidade 3: Ferramenta de estudo de similaridade

A ferramenta de estudo de similaridade permite ao usuário verificar a relação de similaridade na estrutura sequencial entre proteínas de um mesmo projeto. Essa ferramenta gera um arquivo no formato de entrada para o *software Cytoscape*, o qual gera grafos de fácil interpretação. Essa ferramenta funciona da seguinte forma, o usuário seta um valor de similaridade e a ferramenta seleciona somente as proteínas que compartilham desse valor. Assim, esses dados são preenchidos em um arquivo no formato de entrada do *software Cytoscape*. Um exemplo desta funcionalidade é apresentada no capítulo do Estudo de Caso.

5.3.4 Funcionalidade 4: Alinhamento estrutural

Para qualquer conjunto de moléculas que o pesquisador deseje realizar experimento de triagem virtual, é importante que todas estas proteínas se localizem no mesmo plano de coordenadas. Dessa forma, o pesquisador configura uma única área para o ligante atracar nas moléculas. O funcionamento do alinhamento estrutural na ferramenta *Cat-p-data* é dado da seguinte forma: o usuário seleciona uma proteína do projeto para servir de base e, assim, todas as outras proteínas são alinhadas estruturalmente a esta. Uma pasta com todos os novos arquivos *PDBs* alinhados é gerada no diretório do programa. O algoritmo utilizado nessa ferramenta para alinhar é o *Superimposer* da biblioteca *Biopython*. Esse método necessita de parâmetros de entrada, os quais são os resíduos inicial e final de um segmento da estrutura sequencial da proteína. Esse segmento de resíduos é obtido no resultado de alinhamento sequencial. Portanto, o alinhamento estrutural utiliza informações obtidas no alinhamento sequencial para otimizar todo o processo, seguindo as regras descritas na seção 5.3.2.

Essa funcionalidade é útil para alinhar todas as estruturas macromoleculares de um mesmo projeto, como por exemplo, o conjunto da β -glicosidase pesquisado neste trabalho. O alinhamento estrutural foi utilizado para que as moléculas fossem alinhadas

em um mesmo plano de coordenadas. Dessa forma, pode-se realizar os experimentos de docagem molecular utilizando como ligantes a glicose e a celobiose nas moléculas da β -glicosidase.

A Figura 26 exemplifica um alinhamento estrutural com algumas proteínas do conjunto da β -glicosidase.



Figura 26: Exemplo de alinhamento estrutural com algumas proteínas do conjunto da β -glicosidase.

De acordo com a Figura 26, que representa um exemplo de alinhamento estrutural, pode-se notar que três proteínas foram alinhadas, utilizando uma delas como base. A proteína de código PDB “1OIF” é representada na cor verde, a proteína “1OIM” é representada na cor azul e a proteína “1OIN” é representada na cor rosa.

5.3.5 Funcionalidade 5: Ferramenta de visualização do ligante

A ferramenta de visualização dos ligantes mostra quais são ligantes pertencentes as estruturas de proteínas. Dada a seleção de uma ou mais estruturas de proteínas, a apresentação dos ligantes na interface deste módulo é distinguida nas cadeias da estrutura selecionada, ou seja, cada cadeia da estrutura apresenta os ligantes. Esta funcionalidade possibilita ao usuário identificar quais são os ligantes pertencentes as proteínas de um projeto. Este módulo lista todas os ligantes de todas as proteínas de um projeto, apresentando a quantidade de ligantes no mesmo. Além disso, é permitido ao usuário selecionar uma ou mais estruturas de proteínas e assim listar somente os ligantes presentes nas estruturas selecionadas.

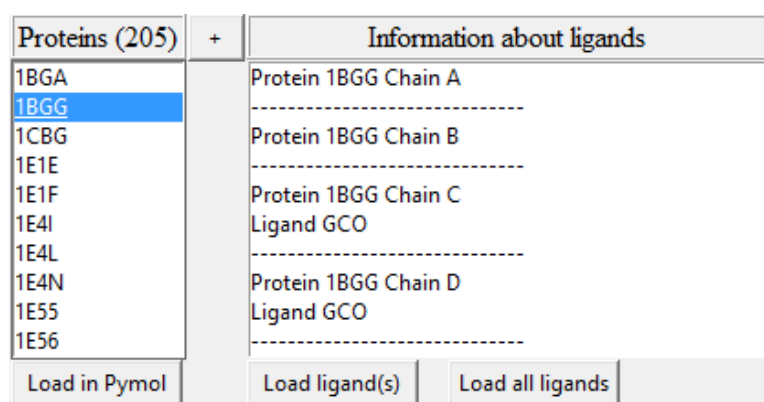


Figura 27: Exemplo do uso da ferramenta de visualização do ligante. Proteína de código PDB 1BGG possui dois ligantes GCO (glicose) nas cadeias C e D de sua estrutura.

A figura 27 mostra um exemplo de uso da ferramenta de visualização do ligante. Para este tipo de visualização, o usuário carrega somente uma proteína de interesse para descobrir os seus ligantes. Como mostra a figura 27, a proteína de código PDB 1BGG possui dois ligantes GCO (glicose) nas cadeias C e D.

5.3.6 Funcionalidade 6: Ambiente tridimensional de visualização de proteínas

Nesta seção é apresentado o ambiente de visualização tridimensional de proteínas, o qual está integrado na ferramenta desenvolvida neste trabalho. Com esse ambiente, o usuário poderá carregar uma ou mais proteínas ao mesmo tempo e dessa forma visualizar as estruturas macromoleculares das mesmas, além de realizar uma série de manipulações como por exemplo, seleção de átomos, seleção de resíduos, selecionar diferentes formas de visualização, podendo ainda salvar imagens das proteínas em diferentes formatos. Com a visualização de várias proteínas ao mesmo tempo, é possível visualizar quais delas estão em um mesmo plano de coordenadas, e assim verificar quais estão preparadas em relação as suas coordenadas para a execução de testes de docagem molecular.

O software *Pymol Viewer* foi o escolhido para integrar um ambiente tridimensional de visualização a ferramenta desenvolvida neste trabalho.

Na figura 28, de exemplo, 3 estruturas com código PDB 1BGA, 1BGG e 1CBG foram selecionadas para a visualização. A ferramenta permite selecionar uma ou mais proteínas ao mesmo tempo. Esse ambiente de visualização é útil também para verificar o resultado do alinhamento estrutural.

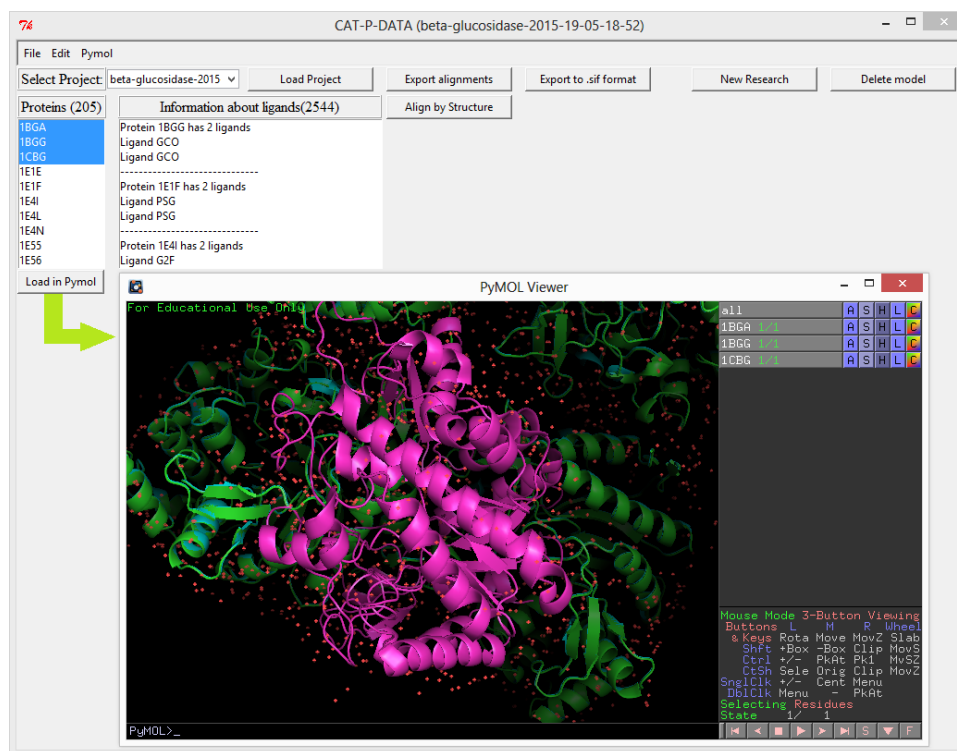


Figura 28: Proteínas do conjunto da β -glicosidase sendo visualizadas no ambiente tridimensional através da nossa ferramenta.

5.3.7 Funcionalidade 7: Módulo de atualização do conjunto de proteínas

O módulo de atualização do conjunto de proteínas tem o objetivo de buscar novas proteínas no servidor do PDB para serem vinculadas aos projetos já existentes do usuário. Este módulo tem seu funcionamento simplificado, pois, utiliza na busca por novas proteínas a mesma palavra-chave utilizada na criação do projeto. A figura 29 apresenta 33 novas proteínas relacionadas à β -glicosidase.

O módulo de atualização é acessado através do menu “Tools” da barra superior da ferramenta Cat-p-Data. Como pode ser visto na Figura 29, o módulo contém uma lista onde as novas proteínas são apresentadas. O botão para inserir as novas proteínas ao projeto é nomeado na interface como “Insert new proteins”. Se o usuário não deseja uma ou mais das novas proteínas apresentadas na lista da interface, basta utilizar o botão “Delete item” para excluir as indesejáveis antes de realizar a aquisição.

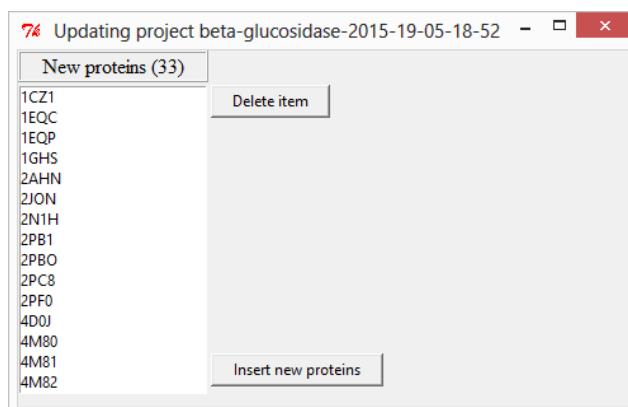


Figura 29: Interface do módulo de atualização utilizando como exemplo novas proteínas da β -glicosidase.

Fonte: Elaborado pelo autor.

5.3.8 Funcionalidade 8: Módulo de limpeza do conjunto de proteínas

O módulo de limpeza do conjunto de proteínas tem o objetivo de eliminar proteínas que não pertencem ao grupo almejado pelo usuário. A busca no servidor do PDB considera sub-palavras da palavra-chave fornecida e, assim, algumas das proteínas adquiridas pelo usuário podem não pertencerem ao grupo desejado. Por exemplo, uma busca na ferramenta que utiliza a palavra-chave “beta-glicosidase”, pode considerar proteínas que possuem somente uma dessas sub-palavras e, assim, listar para aquisição proteínas que não pertencem ao grupo desejado.

Para acessar este módulo, o usuário deve clicar em “Tools” no menu superior do Cat-p-data e após em “Clean project”. A Figura 30 apresenta o módulo de limpeza do conjunto de proteínas utilizando o conjunto de proteínas da β -glicosidase.

O módulo de limpeza do conjunto de proteínas utiliza os resultados do alinhamento sequencial para auxiliar na exclusão de proteínas indesejadas. Como pode ser visto na Figura 30, o processo de limpeza utiliza a estrutura primária de uma proteína do conjunto para servir de base. Após o usuário selecionar a proteína base e clicar no botão “Load similarities”, os resultados de similaridade são apresentados na interface do módulo. Com as informações dessa listagem, o usuário poderá visualizar o nível de limpeza que realizará no projeto. Para iniciar a limpeza, o usuário deve clicar no botão “Delete structures” da interface e, assim, uma segunda tela é renderizada. Nessa segunda tela, o usuário deve preencher dois campos, o primeiro indica um operador matemático, que pode ser o símbolo “>”, “<”, “>=”, “<=” ou “=” e o segundo campo indica o percentual. A ferramenta tem o valor padrão de “<= 80”, o que significa que permanecerá no conjunto somente as proteínas que possuem mais de 80% de similaridade com a proteína base selecionada anteriormente.

Na Figura 30 de exemplo, dado que o usuário escolheu como filtro de exclusão a similaridade de menor ou igual a 80%, as proteínas de código PDB 1QOX(68%), 3TA9(54%),

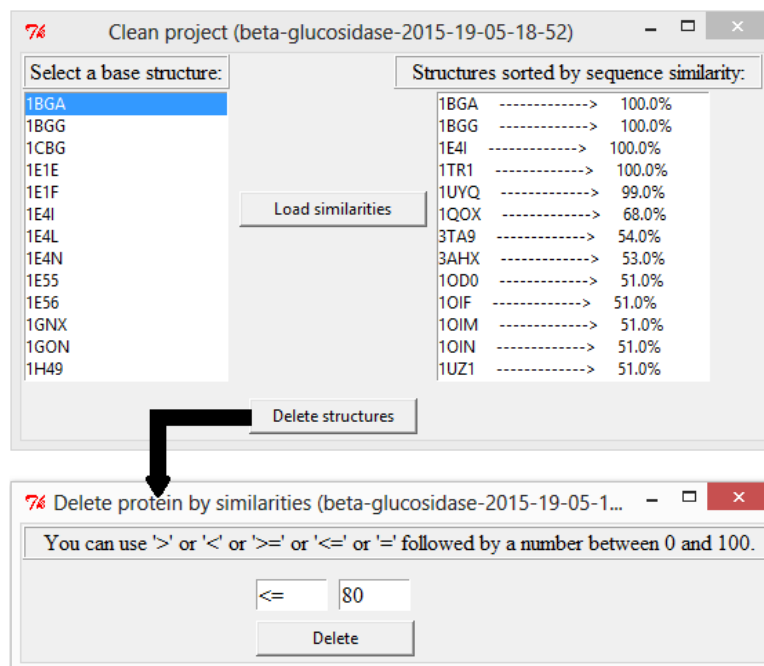


Figura 30: Interface do módulo de limpeza do conjunto de proteínas utilizando como exemplo o conjunto de proteínas da β -glicosidase.

Fonte: Elaborado pelo autor.

3AHX(53%), 1OD0(51%), 1OIF(51%), 1OIM(51%), 1OIN(51%), 1UZ1(51%) e outras com similaridade abaixo de 80% são excluídas do projeto. Nesse exemplo, as proteínas que permanecem no conjunto são as de código PDB 1BGA(100%), 1BGG(100%), 1E4I(100%), 1TR1(100%) e a 1UYQ(99%).

5.3.9 Funcionalidade 9: Módulo de edição de conjunto de proteínas

O módulo de edição de conjunto de proteínas tem o objetivo de vincular proteínas de um projeto qualquer existente na ferramenta ao projeto atual carregado pelo usuário. Este módulo é utilizado quando o usuário deseja realizar experimentos relacionando duas moléculas de conjuntos diferentes. Assim, basta vincular uma ou mais moléculas ao projeto desejado. A figura 31, exemplifica o uso desse módulo vinculando uma nova proteína ao projeto da β -glicosidase.

Como mostrado na figura 31, o exemplo trata da vinculação da proteína “2KSM”, a qual pertence originalmente ao conjunto de proteínas da tuberculose, ao grupo da β -glicosidase. Neste módulo, o usuário seleciona uma ou mais proteínas na lista à esquerda da interface e, após isso, clica no botão “Bind protein(s) from to” no centro da tela. Dessa forma as proteínas são vinculadas a um novo projeto. Importante ressaltar que as moléculas vinculadas a um novo projeto não são desvinculadas do seu projeto original.

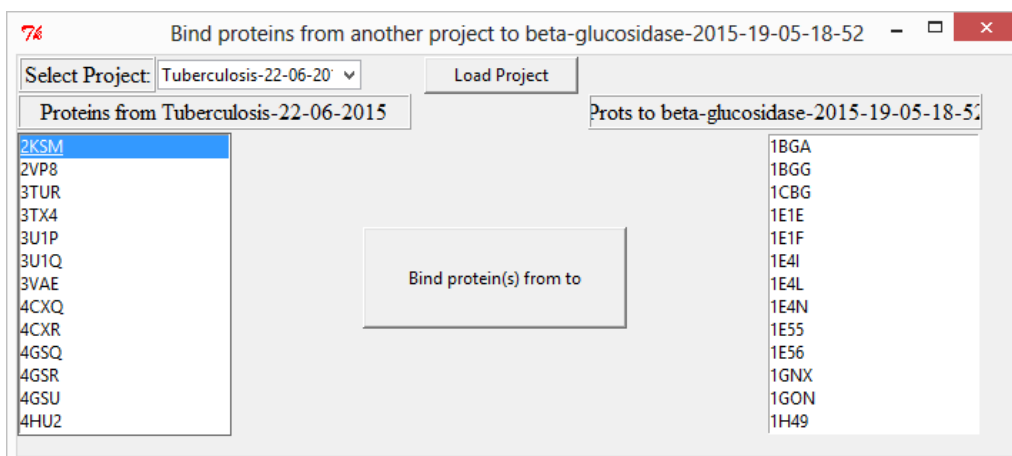


Figura 31: Exemplo do uso do módulo de edição do conjunto de proteínas.

Fonte: Elaborado pelo autor.

5.3.10 Funcionalidade 10: Módulo de inserção de novas proteínas

O módulo de inserção de novas proteínas tem o objetivo de vincular novas proteínas encontradas no servidor do PDB a um projeto específico já existente na ferramenta Cat-p-data. Para acessar esse módulo, o usuário deve clicar em “Tools” no menu superior do Cat-p-data e depois em “Insert new proteins”. Dessa forma, a tela mostrada na Figura 32 aparecerá.

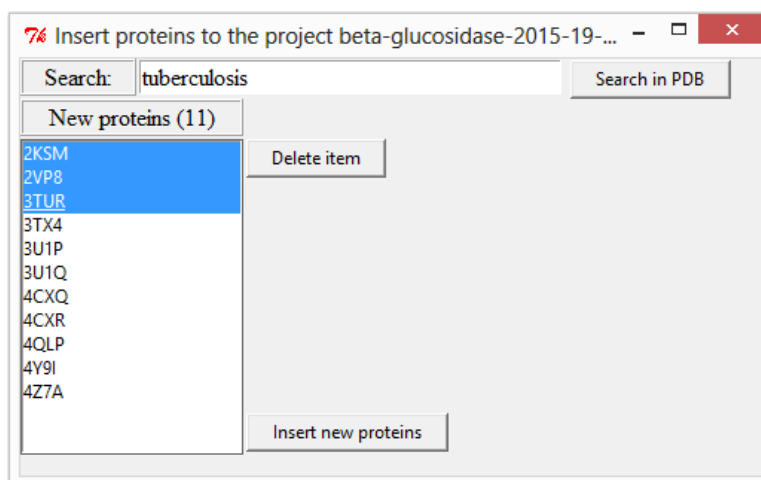


Figura 32: Exemplo do uso do módulo de inserção de novas proteínas em um conjunto de proteínas já existente no Cat-p-Data.

Fonte: Elaborado pelo autor.

A Figura 32, apresenta um exemplo do uso do módulo de inserção de novas proteínas. Como pode ser visto na Figura 32, este módulo possui um campo de busca para o usuário pesquisar por proteínas através de uma palavra-chave. Após a busca, as proteínas são listadas e podem ser adquiridas pelo usuário. É importante notar que todas as proteínas que permanecerem na lista serão inseridas no conjunto atual carregado.

No exemplo da Figura 32, as proteínas da tuberculose que estão marcadas com a cor

azul foram selecionadas para serem excluídas e, assim, somente o restante das moléculas serão inseridas no projeto. Essa funcionalidade ajuda na composição de um conjunto diversificado de moléculas.

6 RESULTADO 2: ESTUDO DE CASO β -GLICOSIDASE

6.1 β -glicosidases

O objetivo principal deste trabalho consiste na implementação de uma ferramenta que possibilite o usuário a investigar grupos específicos de estruturas macromoleculares diversas, como por exemplo, a enzima da β -glicosidase, a qual sua estrutura molecular é investigada nesse trabalho com o objetivo de entender o seu funcionamento em relação ao processo de quebra de açúcares fermentáveis.

Em geral os biocombustíveis podem ser considerados como fontes energéticas infinitas e aparecem como uma alternativa significativa aos combustíveis fósseis. Nesse contexto, evidencia-se que a otimização do processo de geração de açúcares fermentáveis para a produção de bioetanol torna-se um fator importante (GOTTSCHALK; OLIVEIRA; BON, 2010). Segundo (GOTTSCHALK; OLIVEIRA; BON, 2010), as paredes celulares das plantas são uma considerável fonte lignocelulósica para a extração de açúcares fermentáveis (fontes energéticas para a produção de bioetanol). Segundo (LYND et al., 2002), durante a hidrólise enzimática da celulose por micro-organismos (bactérias e fungos) são necessários três tipos de celulasas, as endo-glucanases, as exo-glucanases (celodextrinase e celobiohidrolase) e as β -glicosidases. As β -glicosidases agem sobre substratos solúveis (celobiose) e assim podem ser estudadas pelos modelos tradicionais de cinética enzimática (JEOH et al., 2005). A celobiose ou dissacarídeo representa um importante fator em relação a eficiência do sistema de degradação de matéria lignocelulósica, pois o aumento da concentração de celobiose faz com que aconteça a inibição das outras celulasas (endo-glucanases e exo-glucanases) envolvidas no processo. De acordo com (GOTTSCHALK; OLIVEIRA; BON, 2010), a β -glicosidase é responsável pela última etapa do processo de hidrólise lignocelulósica que consiste em quebrar a celobiose (dissacarídeo) em duas moléculas (monossacarídeos) que representam o produto final (glicose). Porém, outro problema surge quando acontece o acúmulo de glicose (moléculas monossacarídeas), que resulta na inibição da enzima β -glicosidase e acarreta em uma menor taxa de extração de açúcares fermentáveis (glicose). Dessa forma, pretende-se neste trabalho analisar e investigar *in-silico* sobre o comportamento da β -glicosidase no pro-

cesso de hidrólise através de criação de mutações desta enzima e também testes de docagem molecular com o objetivo de encontrar melhores sítios de ligação para a celobiose, a qual é a molécula que a β -glicosidase interage na hidrólise.

6.2 Estudo de caso

O estudo de caso realizado com as proteínas da β -glicosidase foi executado da seguinte forma: primeiramente foram adquiridas as proteínas do grupo da β -glicosidase através da ferramenta *Cat-p-data* utilizando no campo de busca a palavra-chave “beta-glucosidase”. As 205 proteínas encontradas foram armazenadas no banco de dados local da ferramenta e vinculadas a um projeto. Com as moléculas inseridas no banco de dados, a funcionalidade de alinhamento sequencial foi utilizada. Todas as proteínas desse conjunto foram alinhadas entre si e a informação obtida foi armazenada no banco de dados.

A Figura 33 apresenta o formato do alinhamento sequencial para o conjunto de proteínas da β -glicosidase adquiridos com a ferramenta *Cat-p-Data*.

	1BGA	1BGG	1CBG	1E1E	1E1F	1E4I	1E4L	1E4N	1E55	1E56	1GNX	1GON	1H49	1H4P
1BGA	100	100	45	43	43	100	43	43	43	43	48	48	43	34
1BGG	100	100	45	44	44	99	43	43	43	43	48	48	43	34
1CBG	45	45	100	51	51	44	51	51	51	51	45	45	51	33
1E1E	43	44	51	100	100	43	100	100	100	100	43	43	100	32
1E1F	43	44	51	100	100	43	100	100	100	100	43	43	100	32
1E4I	100	99	44	43	43	100	43	43	43	43	48	48	43	34
1E4L	43	43	51	100	100	43	100	100	100	100	43	43	100	32
1E4N	43	43	51	100	100	43	100	100	100	100	43	43	100	32
1E55	43	43	51	100	100	43	100	100	100	100	43	43	100	32
1E56	43	43	51	100	100	43	100	100	100	100	43	43	100	32
1GNX	48	48	45	43	43	48	43	43	43	43	100	100	43	33
1GON	48	48	45	43	43	48	43	43	43	43	100	100	43	33
1H49	43	43	51	100	100	43	100	100	100	100	43	43	100	32
1H4P	34	34	33	32	32	34	32	32	32	32	33	33	32	100

Figura 33: Parcela do alinhamento sequencial do conjunto de proteínas da β -glicosidase.

Os valores presentes no alinhamento sequencial apresentados na Figura 33 são percentuais. Pode-se ver que a proteína de código PDB:1BGA possui 100% de identidade com a proteína de código PDB:1BGG. Com essas informações, pode-se descobrir proteínas que compartilham características importantes.

A Figura 34 apresenta um grupo de diferentes β -glicosidases que possuem 90% ou mais de similaridade em sua estrutura primária. Através do uso da ferramenta de estudo de sequência implementado no *Cat-p-Data*, foi possível extrair esse tipo de informação. Essa ferramenta funciona da seguinte forma: o usuário informa o valor de similaridade desejado para formar os conjuntos de proteínas e um arquivo com as relações entre proteínas é gerado. Como pode ser visto na Figura 34, para o estudo de caso da β -glicosidases foi utilizado 90% de similaridade, o que resultou em diversos grupos de proteínas compartilhando essa similaridade. A Figura 34 apresenta somente o maior grupo gerado, possuindo 26 proteínas. Com esses grupos, é possível ao pesquisador descobrir conjuntos de proteínas que possuem diferentes níveis de similaridade.

Um dos objetivos deste trabalho refere-se à pesquisa em relação à interação da enzima β -glicosidase com o dissacarídeo “celobiose” e o monossacarídeo “glicose”. Como parte

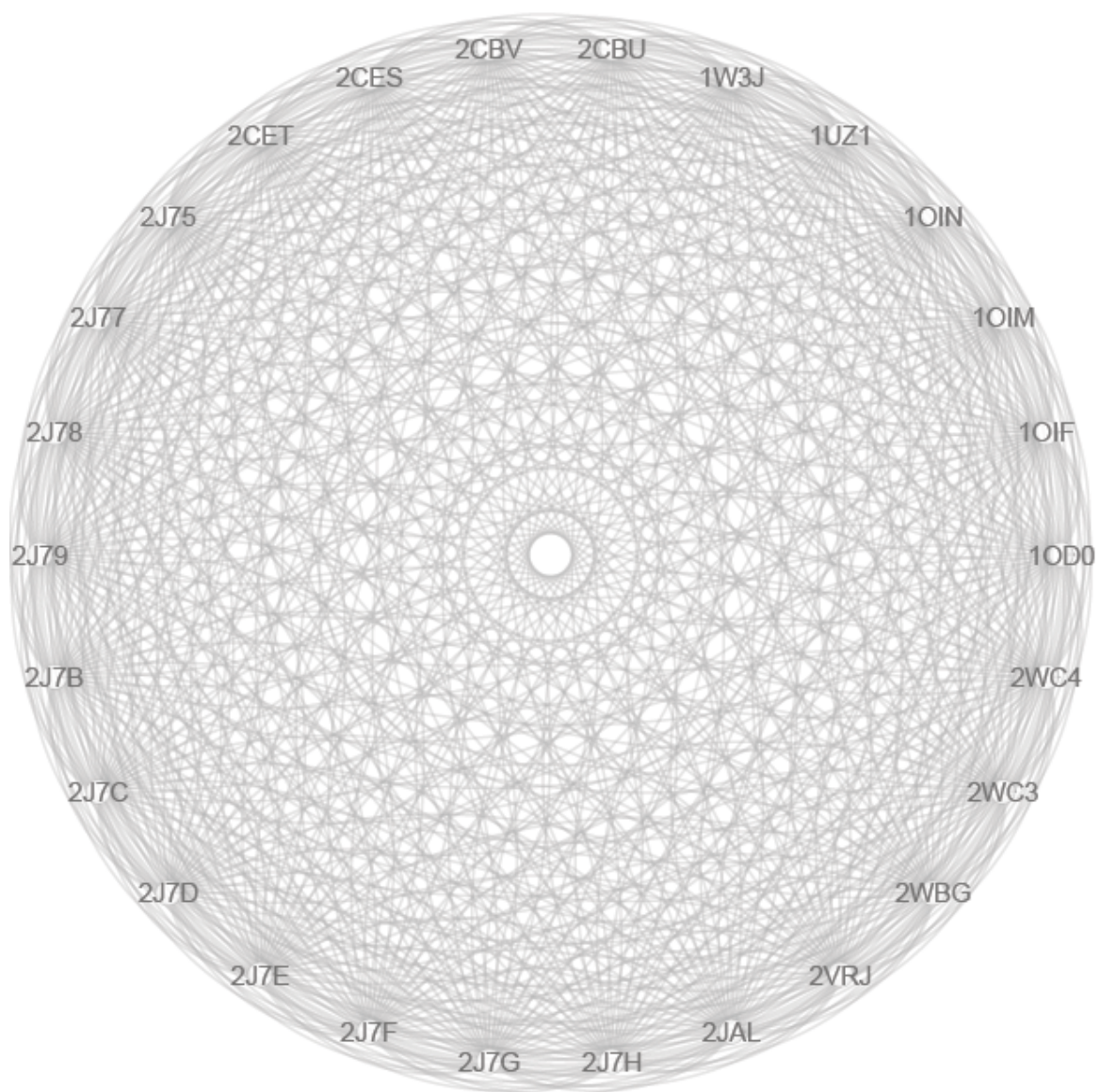


Figura 34: Grupo de β -glicosidases com 90% ou mais de similaridade.
Fonte: Imagem gerada por meio do *software* Cytoscape

do trabalho, tem por objetivo investigar o sítio ativo da β -glicosidase, para assim, entender a interação desta enzima com “celobiose” e “glicose”. Segundo (GOTTSCALK; OLIVEIRA; BON, 2010) a glicose é o produto final no processo de extração de açúcares fermentáveis do bagaço da cana-de-açúcar, utilizado para a produção de bio-etanol. Considerando que o experimento de VS possui diferentes estruturas de proteínas da β -glicosidase e dois ligantes, surge a necessidade de alinhar todas essas estruturas em um mesmo plano de coordenadas para facilitar a configuração desse experimento. Assim, a funcionalidade de alinhamento estrutural implementada foi utilizada, o que auxiliou e acelerou a preparação das proteínas para os experimentos de VS. Para alcançar melhores resultados no alinhamento estrutural, esse módulo foi construído utilizando o resultado gerado pelo alinhamento sequencial juntamente com a criação de heurísticas, as quais foram discutidas na 5.3.2.

Utilizando a ferramenta de alinhamento estrutural, todas as 26 estruturas de proteínas apresentadas na Figura 34 foram alinhadas. Após isso, considerando essas 26 estruturas, experimentos de docagem molecular foram realizados para avaliação da interação proteína-ligante (β -glicosidase-glicose e β -glicosidase-celobiose). A Figura 35 apresenta a estrutura terciária de uma das β -glicosidases utilizadas nesse experimento e a Figura 36 apresenta a estrutura molecular dos ligantes glicose e celobiose.

O resultado dos experimentos de VS considerando as 26 estruturas de proteínas da β -glicosidase e os ligantes glicose e celobiose são apresentados nos gráficos das Figuras 37 e 38.

Como pode ser visto nos gráficos das Figuras 37 e 38, o eixo-y apresenta os valores de *FEB* para as proteínas do eixo-x. Com relação a interação do ligante glicose com as β -glicosidases, o melhor resultado em relação à *FEB* foi com a molécula de código PDB:2CBV que apresentou um valor de -6.4 Kcal/mol. Utilizando a celobiose como ligante, o melhor resultado foi de -7.4 Kcal/mol para a *FEB*, porém nesse caso, 5 moléculas diferentes obtiveram esse valor na interação com o ligante. Essas moléculas foram as de código PDB: 1W3J, 2J78, 2J7C, 2J7G, 2WBG. Também foram executadas simulações de docagem molecular com outras estruturas, porém, os resultados não serão discutidos neste trabalho.

Após o experimento de VS ter sido efetuado, muitos dados são gerados, nesse de estudo de caso foram 26 experimentos de docagem molecular com o ligante glicose e 26 experimentos com o ligante da celobiose. Para analisar todas esses resultados, um *script* para a análise das interações molécula-ligante gerado pela ferramenta discutida na seção 3.2.2, foi executado. Com esse *script*, foram gerados dados e gráficos que facilitam a visualização das interações entre a molécula-alvo e o ligante para cada experimento de docagem molecular executado. O *script* utiliza a ferramenta *Ligplot* para geração das imagens no formato *eps*. As figuras 39 e 40 apresentam os melhores resultados de interação molecular nesse experimento. Cada resultado de interação molecular destaca os



Figura 35: β -glicosidase de código PDB:2CBU.

Fonte: Imagem gerada por meio do *software Pymol*

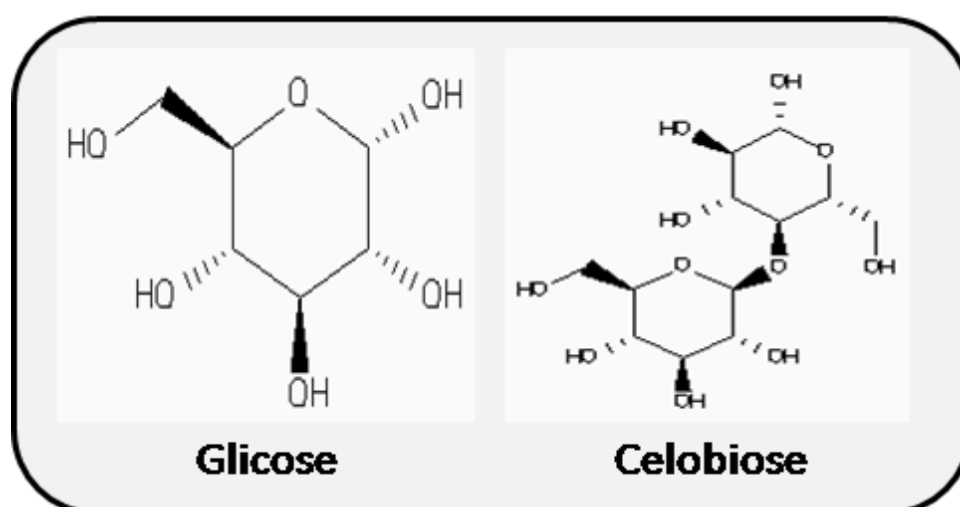


Figura 36: Estrutura molecular dos ligantes glicose e celbiose utilizados no experimento de VS do estudo de caso.

Fonte: Elaborada pelo autor.

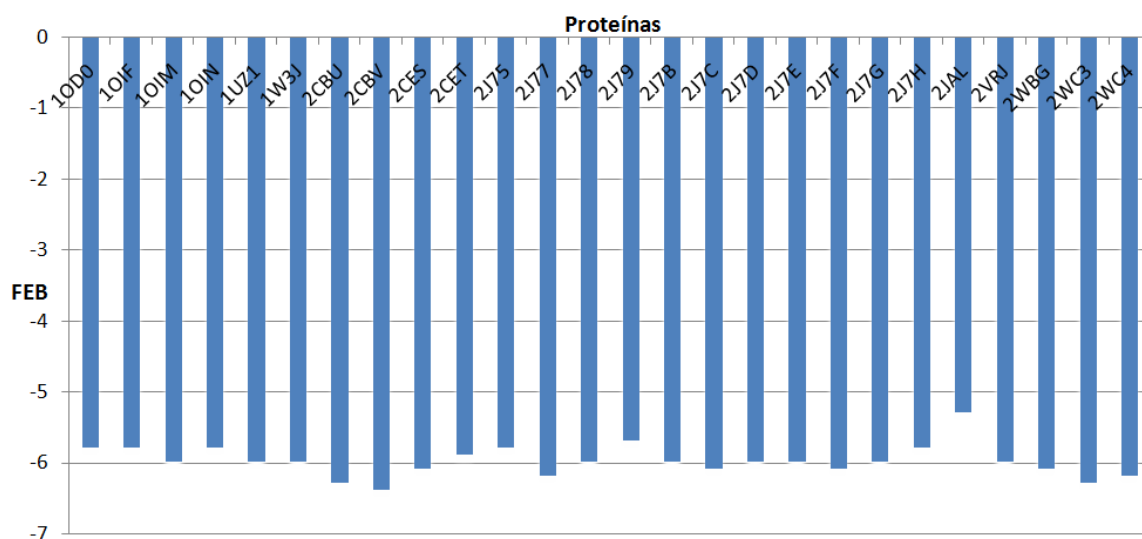


Figura 37: Resultado do experimento de VS utilizando a glicose como ligante e as β -glicosidases como moléculas-alvo.

Fonte: Elaborada pelo autor.

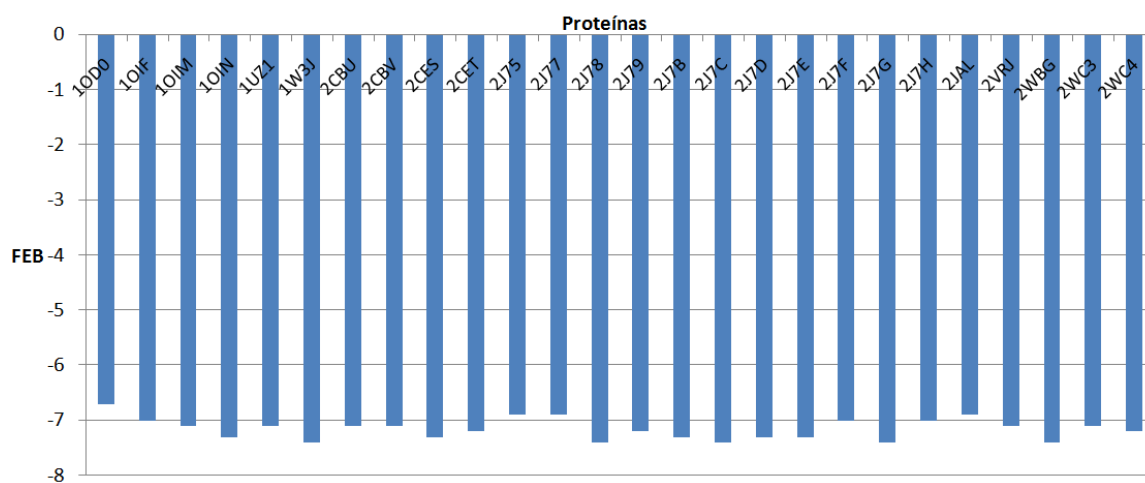


Figura 38: Resultado do experimento de VS utilizando a celobiose como ligante e as β -glicosidases como moléculas-alvo

Fonte: Elaborada pelo autor.

principais aminoácidos que fizeram contato com os átomos dos ligantes.

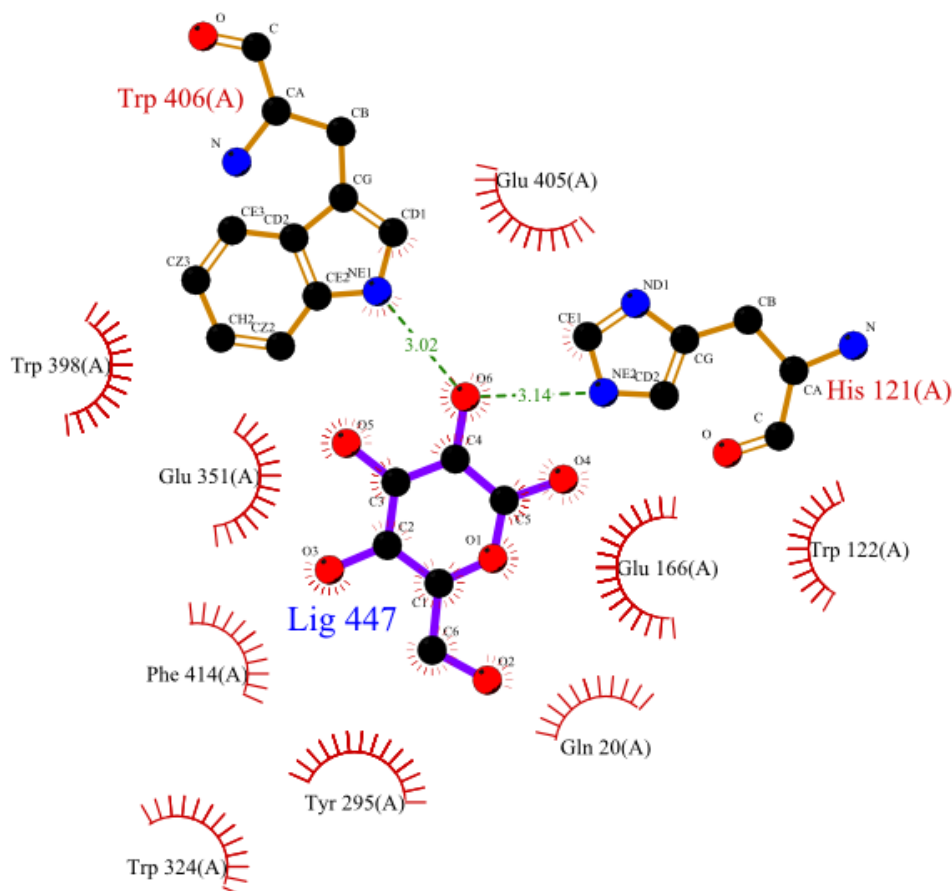


Figura 39: Interação molecular entre os resíduos da proteína de código PDB:2CBV e os átomos do ligante glicose.

Fonte: Gerado pelo *software* Ligplot.

Outra ferramenta utilizada na análise dos resultados de interação proteína-ligante da β -glicosidase foi o *software* nAPOLI. O nAPOLI verifica as interações entre os resíduos da molécula-alvo e os átomos do ligante e apresenta todos esses resultados em uma interface *web* de fácil entendimento.

Nesse estudo de caso, os resultados de triagem virtual da β -glicosidase com os ligantes glicose e celobiose foram utilizados como entrada para a ferramenta nAPOLI e, assim, as interações proteína-ligante foram analisadas. Com isso, pode-se descobrir resíduos da molécula da β -glicosidase que interagiram com a glicose e celobiose e resíduos que interagiram com somente um dos ligantes. Nessa pesquisa, supõe-se que aminoácidos importantes são aqueles que fazem contato com a glicose e não fazem com a celobiose. Considerando isso, com relação a β -glicosidase, aminoácidos candidatos à mutações pontuais são aqueles que interagem com a glicose e não interagem com a celobiose, pois, espera-se que, após a mutação pontual desses aminoácidos, a enzima da β -glicosidase mude seu comportamento e passe a não interagir mais com a glicose e somente interaja com a celobiose.

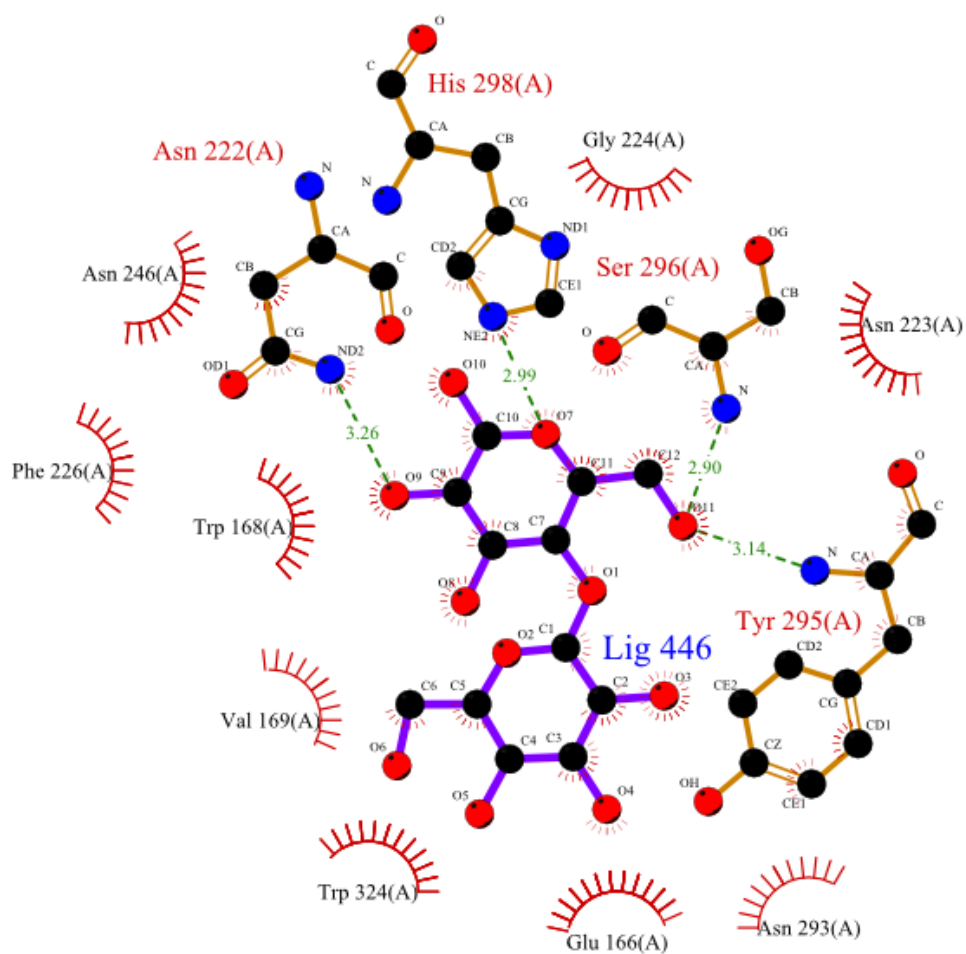


Figura 40: Interação molecular entre os resíduos da proteína de código PDB:2WBG e os átomos do ligante celobiose.

Fonte: Gerado pelo *software Ligplot*.

Tabela 3: Top 8 dos aminoácidos em contato com glicose e/ou celobiose durante os experimentos.

Aminoácidos em contato com Glicose e/ou Celobiose	% contatos Glicose	% contatos Celobiose
GLU166	53%	100%
TRY295	100%	100%
ASN222	34%	100%
ASN223	34%	100%
TRP324	61%	100%
TRP122	65%	84%
TRP406	65%	73%

A tabela 3 mostra os resultados das interações das β -glicosidases com os ligantes glicose e celobiose nos experimentos. Assim, pode-se observar que a maioria dos aminoácidos das β -glicosidases fizeram contatos com a glicose e celobiose, que é o caso dos aminoácidos GLU166, TRY295, ASN222, ASN223, TRP324, TRP122 e TRP406.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

A aquisição de arquivos de proteínas do PDB seguido da análise desses dados é uma tarefa que pode ser custosa ao usuário se não for realizada por uma ferramenta computacional. De forma manual, após a aquisição dos arquivos PDB, o usuário terá estes dados armazenados em um diretório do seu próprio computador e as análises terão que ser realizadas por diferentes ferramentas. Tendo em vista o grande crescimento das bases de dados proteicas e a dificuldade de caracterização de conjuntos de proteínas, acredita-se que a ferramenta desenvolvida neste trabalho ajuda na aquisição de conjuntos específicos de moléculas interessantes a uma determinada pesquisa, além de facilitar a caracterização desses conjuntos de proteínas através dos diferentes módulos implementados e integrados em um só ambiente.

O estudo de caso com a enzima da β -glicosidase mostra que utilizando a ferramenta *Cat-p-Data* é possível adquirir um conjunto específico de proteínas, e a partir disso, filtrar e extrair grupos de moléculas importantes que possuam determinada similaridade nas suas estruturas. A partir desse estudo de caso, pode-se ver que a ferramenta é genérica, permitindo o pesquisador trabalhar com qualquer conjunto de proteínas. A partir dos resultados do estudo de caso importantes características do modo de ligação da β -glicosidase com a glicose e celobiose foram extraídas.

As próximas etapas desse trabalho são relacionadas à implementações de novas funcionalidades para facilitar as pesquisas sobre proteínas. Entre as novas funcionalidades, estão as seguintes:

- Integração do módulo de mineração de dados em desenvolvimento em outro trabalho;
- Integração do módulo de docagem molecular em desenvolvimento em outro trabalho;
- Integração do módulo de interação proteína-ligante em desenvolvimento em outro trabalho;
- Busca e estudo de novos métodos para verificação da similaridade entre proteínas;

- Disponibilizar a ferramenta proposta na página do grupo;
- Criação da documentação da ferramenta;
- Teste com usuários - análise do uso da interface.

REFERÊNCIAS

ARNDT, T. Visual software tools for bioinformatics. **Journal of Visual Languages and Computing**, p.291–301, 2008.

AUERBACH, C. **Mutation research**: problems, results and perspectives. Springer, 2013.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. **Biochemistry. 5th**. New York: WH Freeman, 2002.

BERMAN, H.; HENRICK, K.; NAKAMURA, H.; MARKLEY, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. **Nucleic acids research**, v.35, n.suppl 1, p.D301–D303, 2007.

BERMAN, H.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T.; WEISSIG, H.; SHINDYALOV, I.; BOURNE, P. The protein Data Bank. **Nucleic Acids Research**, v.28, p.235–242, 2000.

CAVASSOTO, C.; ORRY, A. Ligand docking and structure-based virtual screening in drug discovery. **Current Topics in Medical Chemistry**, v.7, p.1006–14, 2007.

COCK, P. J. A.; ANTAO, T.; CHANG, J. T.; CHAPMAN, A. B.; COX, C. J.; DALKE, A.; FRIEDBERG, I. HAMELRYCK, T.; KAUFF F. WILCZYNSKI, B.; HOON, M. J. L. de. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v.25, n.11, p.1422–1423, 2009.

COHEN, J. Bioinformatics — An Introduction for Computer Scientists. **ACM Computing Surveys**, v.36, n.2, p.122–158, 2004.

COORDINATORS, N. R. Database resources of the national center for biotechnology information. **Nucleic acids research**, v.41, n.Database issue, p.D8, 2013.

COSCONATI, S.; FORLI, S.; PERRYMAN, A.; HARRIS, R.; GOODSSELL, D. S.; OLSON, A. Virtual Screening with AutoDock: Theory and Practice. **NIH Public Access**, v.5, p.597–607, 2010.

FERREIRA, L. G.; SANTOS, R. N. dos; OLIVA, G.; ANDRICOPULO, A. D. Molecular Docking and Structure-Based Drug Design Strategies. **Molecules**, v.20, n.7, p.13384, 2015.

FEYFANT, E.; SALI, A.; FISER, A. Modeling mutations in protein structures. **Protein Science**, v.16, p.2030–2041, 2007.

GONÇALVES, R. S. W.; GONÇALVES-ALMEIDA, M. V.; ARRUDA, L. A.; MEIRA JR, W.; SILVEIRA, H. C. da; D., P. V. E.; MELO-MINARDI, C. R. de. PDBest: a user-friendly platform for manipulating and enhancing protein structures. **Bioinformatics**, 2015.

GOTTSCHALK, L.; OLIVEIRA, R.; BON, E. Cellulases, xylanases, B-glucosidase and ferulic acid esterase produced by *Trichoderma* and *Aspergillus* act synergistically in the hydrolysis of sugarcane bagasse. **Biochemical Engineering**, v.51, p.72–78, 2010.

IRWIN, J.; STERLING, T.; MYSINGER, M.; BOLSTAD, E.; COLEMAN, R. ZINC: A free tool to discovery chemistry for biology. **Journal of Computational Chemistry**, p.1757–1768, 2012.

JEON, T.; BAKER, J.; ALI, M.; HIMMEL, M.; ADNEY, W. B-d-glucosidase reaction kinetics from isothermal titration microcalorimetry. **Analytical biochemistry**, v.347, p.244–253, 2005.

KAWASHIMA, S.; POKAROWSKI, P.; POKAROWSKA, M.; KOLINSKI, A.; KATAYAMA, T.; KANEHISA, M. AAindex: amino acid index database, progress report 2008. **Nucleic Acids Research**, v.36, 2008.

KUNTZ, I. D. Structure-based Strategies for Drug Design and Discovery. **Science**, v.257, p.1078–1082, 1992.

LESK, M. L. **Introduction to Bioinformatics**. 2002. v.1.

LOUNNAS, V.; RITSCHER, T.; KELLER, J.; MCGUIRE, R.; BYWATER, R. P.; FOLLOPE, N. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. **Computational and structural biotechnology journal**, v.5, n.6, p.1–14, 2013.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? An introduction and overview. **Yearbook of Medical Informatics**, p.83–100, 2001.

LYND, L.; WEIMER, P.; ZYL, W.; PRETORIUS, I. Microbial cellulose utilization: fundamentals and biotechnology. **Microbiology and Molecular Biology Reviews**, v.66, p.506–542, 2002.

MAVROMOUSTAKOS, T.; DURDAGI, S.; KOUKOULITSA, C.; SIMCIC, M.; G PA-
PADOPOULOS, M.; HODOSCEK, M.; GOLIC GRDADOLNIK, S. Strategies in the
rational drug design. **Current medicinal chemistry**, v.18, n.17, p.2517–2530, 2011.

MCINNES, C. Virtual screening strategies in drug discovery. **Current opinion in chemi-
cal biology**, v.11, n.5, p.494–502, 2007.

MENG, X. Y. Y.; ZHANG, H. X. X.; CUI, M. Molecular docking: a powerful approach
for structure-based drug discovery. **Current computer-aided drug design**, v.7, p.146–
157, 2011.

MORRIS, G. M.; HUEY, R.; LINDSTROM, W.; SANNER, M. F.; BELEW, R. K.; GO-
ODSELL, D. S.; OLSON, A. J. Autodock4 and AutoDockTools4: automated docking
with selective receptor flexibility. **Computational Chemistry**, v.16, p.85–91, 2009.

OJO, O. O.; OMABE, M. Incorporating bioinformatics into biological science education
in Nigeria: Prospects and challenges. **Elsevier**, p.784–787, 2011.

PRLIĆ, A.; BLIVEN, S.; ROSE, W. P.; BLUHM, F. W.; BIZON, C.; GODZIK, A.;
BOURNE, E. P. Pre-calculated protein structure alignments at the RCSB PDB website.
Bioinformatics, 2010.

RUIZ-BLANCO, Y. B.; PAZ, W.; GREEN, J.; MARRETO-PONCE, Y. ProtDCal: A pro-
gram to compute general-purpose numerical descriptors for sequences and 3D-structures
of proteins. **Bioinformatics**, v.16, p.162, 2015.

SCHRÖDINGER, L. The PyMOL Molecular Graphics System. , 2010.

SEUS, V. R.; SILVA, L.; GOMES, J.; SILVA, P. E. A. da; WERHLI, A. V.; MACHADO,
K. S.; PRATES, N. S.; ZANATTA, N. A Framework for Virtual Screening. In: SYMPO-
SIUM ON APPLIED COMPUTING, 2016, New York, NY, USA. **Anais...** ACM, 2016.

SILVA, J. J.; SILVA, J. A. L. d. D-amino acids in biology - more than one thinks. **Química
Nova**, v.32, n.2, p.554–561, 2009.

SZABADKA, Z.; GROLMUSZ, V. Building a Structured PDB: The RS-PDB Database.
Engineering in Medicine and Biology Society, 2006.

TROTT, O.; OLSON, A. AutoDock Vina: improving the speed and accuracy of docking
with a new scoring function, efficient optimization and multithreading. **Journal of Com-
putational Chemistry**, v.31, p.455–461, 2010.

VERLI, H. **Bioinformática: da Biologia à Flexibilidade Molecular**.
<http://www.ufrgs.br/bioinfo/ebook>: [s.n.], 2014. v.1.

WALLACE, A. C.; LASKOWSKI, R. A.; THORNTON, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. . **Protein Eng.**, v.8, p.127–134, 1996.

WANG, R.; FANG, X.; L., Y.; YANG, C.; WANG, S. The PDBbind Database: Methodologies and Updates. **Journal of Medical Chemistry**, v.48, p.4111–4119, 2005.

WINCK, A.; MACHADO, K.; SOUZA, O. Norberto-de; RUIZ, D. FReDD: Supporting Mining Strategies through a Flexible-Receptor Docking Database. **Advances in Bioinformatics and Computational Biology**, v.5676, p.143–146, 2009.

WINCK, A. T.; MACHADO, K. S.; SOUZA, O. Norberto-de; RUIZ, D. D. A. Supporting Intermolecular Interaction Analyses of Flexible-Receptor Docking Simulations. . **IADIS International Conference Applied Computing, 2010, Timisoara. Applied Computing.**, p.183–190, 2010.

ZHAO, W.; HEVENER, K. E.; WHITE, S. W.; LEE, R. E.; BOYETT, J. M. A statistical framework to evaluate virtual screening. **BMC bioinformatics**, v.10, n.1, p.1, 2009.