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

EN-MUTATE: predição do impacto de mutações pontuais em proteínas utilizando *Ensemble Learning*

Alex Dias Camargo

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 dos Santos Machado
Coorientador: Prof. Dr. Adriano Velasque Werhli

Rio Grande, junho de 2017

Ficha catalográfica

C172e Camargo, Alex Dias.
EN-MUTATE: predição do impacto de mutações pontuais em proteínas utilizando *Ensemble Learning* / Alex Dias Camargo. – 2017.
101 p.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-graduação em Engenharia de Computação, Rio Grande/RS, 2017.
Orientadora: Drª. Karina dos Santos Machado.
Coorientador: Dr. Adriano Velasque Werhli.

1. Mutações pontuais 2. Predição de estabilidade 3. *Ensemble learnin*g I. Machado, Karina dos Santos II. Werhli, Adriano Velasque III. Título.

CDU 004:57

Catalogação na Fonte: Bibliotecário Me. João Paulo Borges da Silveira CRB 10/2130

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

EN-MUTATE: predição do impacto de mutações pontuais em proteínas utilizando *Ensemble Learning*

Alex Dias Camargo

Banca examinadora:

Prof. Dr. Eduardo Nunes Borges

Profa. Dra. Mariana Recamonde Mendoza

Prof. Dr. Adriano Velasque Werhli
Coorientador

Profa. Dra. Karina dos Santos Machado
Orientadora

AGRADECIMENTOS

Este trabalho foi apoiado pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Edital Biologia Computacional de número 051/2013.

SUMÁRIO

LISTA DE FIGURAS

RESUMO

CAMARGO, Alex Dias. **EN-MUTATE: predição do impacto de mutações pontuais em proteínas utilizando *Ensemble Learning***. 2017. 101 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

A metodologia abordada nesta dissertação é baseada na combinação dos resultados de diferentes ferramentas de predição do impacto de mutações pontuais em proteínas, assumindo-se o pressuposto de *Ensemble Learning* na qual a capacidade de generalização de um conjunto é frequentemente mais forte do que uma decisão individual. O objetivo é predizer qual o impacto que uma mutação pode resultar em um mutante "*in-silico*". Para isso, foram adotadas ferramentas descritas na literatura como capazes de predizer os efeitos na estabilidade de uma proteína sobre mutações pontuais através da variação da energia livre $\Delta\Delta G$, ou seja, a diferença de energia livre entre uma proteína do tipo selvagem e o seu mutante. As primeiras versões da metodologia proposta, EN-MUTATE, realizaram o *ensemble* por meio de uma votação por pluralidade entre as ferramentas integradas. À vista disso, com a necessidade de se expandir as análises com o intuito de permitir uma metodologia baseada em modelos treinados através de diferentes classificadores, a abordagem proposta foi reestruturada e passou a abordar múltiplas opções de predição *ensemble*, o que acabou sendo agregado a ferramenta desenvolvida EN-MUTATE*web*. Um fator relevante a ser mencionado sobre a viabilidade da sua utilização é a dificuldade de seleção de um determinado método *a priori*, tendo em vista que não há como se prever àquele que terá melhor desempenho para os dados de interesse. Do mesmo modo, o trabalho necessário para teste e comparação de múltiplas abordagens pode tornar o tempo de pesquisa demasiadamente alto para o especialista. De forma a mensurar a viabilidade de aplicação de *ensemble learning* ao problema de pesquisa, esta dissertação avaliou seus resultados com base em valores biológicos experimentais, sendo que os experimentos computacionais foram divididos em cinco abordagens com diferentes configurações. Por fim, para os principais conjuntos de dados adotados, a metodologia EN-MUTATE obteve em grande parte modelos mais acurados. Desse modo, as principais contribuições obtidas com o desenvolvimento desta dissertação atendem ao seu principal objetivo: definir uma metodologia cuja finalidade é adotar o conceito de *Ensemble Learning* para combinar em uma única abordagem os resultados de diferentes ferramentas de predição do impacto de mutações pontuais em proteínas, buscando, assim, a adoção de abordagens para produzir um resultado final em conjunto potencialmente melhor do que os individuais.

Palavras-chave: Mutações pontuais, predição de estabilidade, *ensemble learning*.

ABSTRACT

CAMARGO, Alex Dias. **EN-MUTATE: prediction of protein stability changes upon single point mutation using Ensemble Learning**. 2017. 101 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

The methodology used in this dissertation is based on the combination of the results of different tools to predict the impact of point mutations on proteins, using the assumption of Ensemble Learning, in which the capacity of generalization of a set is often stronger than an individual decision. The goal is to predict the impact that a mutation can have on an "in-silico" mutant. To this end, tools described in the literature have been chosen for being capable of predicting the effects on stability of a protein on single point mutations through the free energy variation $\Delta\Delta G$, that is, the free energy difference between a wild-type protein and its mutant. The first versions of the proposed methodology, EN-MUTATE, performed the ensemble by means of a plurality voting among the integrated tools. Having this in view, and considering the need to expand the analyzes in order to allow a methodology based on models trained through different classifiers, the proposed approach was restructured and started to address multiple options of ensemble prediction, which ended up being added to EN-MUTATEweb, the tool developed. A relevant factor to be mentioned about the viability of its use is the difficulty of selecting a certain method *a priori*, considering that there is no way to predict the one that will perform best for the data of interest. Likewise, the work required to test and compare multiple approaches can make the search time too high for the specialist. In order to measure the feasibility of applying ensemble learning to the research problem, this dissertation evaluated its results based on experimental biological values, and the computational experiments were divided into five approaches with different configurations. Finally, in the main datasets used EN-MUTATE methodology obtained in large part more accurate models. Thus, the main contributions of this dissertation cover the main objective: define a methodology whose purpose is to adopt the concept of Ensemble Learning to combine results of different tools to predict the impact of point mutations on proteins, seeking the adoption of approaches to produce a potentially better combined result than the individual ones.

Keywords: Point mutation, stability prediction, ensemble learning.

LISTA DE ABREVIATURAS E SIGLAS

AM - Aprendizado de Máquina

ARFF - *Attribute-Relation File Format*

CUPSAT- *Cologne University Protein Stability Analysis Tool*

DNA - *DeoxyriboNucleic Acid*

FN - Falso negativo

FP - Falso positivo

FURG - Universidade Federal do Rio Grande

IA - Inteligência Artificial

iRDP - *in-silico Rational Designing of Proteins*

KDD - *Knowledge Discovery in Databases*

mCSM - *mutation Cutoff Scanning Matrix*

PDB - *Protein Data Bank*

PGH - Projeto Genoma Humano

RMN - Ressonância Magnética Nuclear

RNA - *RiboNucleic Acid*

SDM - *Site Directed Mutator*

SKEMPI- *Structual database of Kinetics and Energetics of Mutant Protein Interactions*

SMO - *Sequential Minimal Optimization*

SVM - *Support Vector Machine*

VN - Verdadeiro negativo

VP - Verdadeiro positivo

WEKA - *Waikato Environment for Knowledge Analysis*

1 INTRODUÇÃO

A Bioinformática pode ser definida como o emprego de ferramentas computacionais no estudo de problemas e questões biológicas, abrangendo aplicações relacionadas à diversas áreas do conhecimento, como a bioquímica e a ciência da computação (VERLI et al., 2014). O termo Bioinformática foi utilizado inicialmente no ano de 1978 em um artigo publicado por Paulien Hogeweg e Ben Hesper (HOGEWEG, 2011). Desde então, o seu uso se tornou mais abrangente, ganhando popularidade entre 1990 e 2000, impulsionado pelo sucesso do Projeto Genoma Humano (PGH). O PGH tinha como objetivo sequenciar e mapear todos os genes dos seres humanos (PENNISI, 2001).

Ainda assim, bem antes disso, na década de 1930, geneticistas começavam a pesquisar que tipo de proteínas (cadeia linear de aminoácidos covalentes) poderiam apresentar o nível de estabilidade exigido pelos genes e ainda ser capaz de modificar-se de modo estável e repentino, gerando mutantes para sustentar a evolução (WATSON et al., 2015). Segundo Bettelheim et al. (2012), um dos principais aspectos do código genético é a sua universalidade. Em diferentes organismos, de uma bactéria a uma planta, de uma planta ao homem, a mesma sequência de três bases é capaz de codificar um mesmo aminoácido. Com isso, a compreensão do efeito de mutações pontuais em proteínas tornou-se um assunto eminente na biologia molecular, principalmente quando se refere a relação estrutura-função de uma proteína (KHAN; VIHINEN, 2010). Tais adventos, como a predição estrutural de proteínas e o atracamento molecular receptor-ligante, têm sua aplicabilidade desde a análise de dados biológicos até o desenvolvimento de métodos que permitam o uso do computador para tarefas comumente experimentais. Por exemplo, para a indústria farmacêutica, as análises de mutações podem ser usadas para identificar uma série de mudanças que podem ocasionar resistência a determinadas drogas (HARVEY; FERRIER, 2011). Com isso, a acurácia da predição do efeito de mutações pontuais em proteínas é considerada uma importante questão para se aumentar a eficácia experimental de muitas técnicas biológicas (HUANG et al., 2014).

Nas últimas duas décadas, diferentes ferramentas computacionais foram desenvolvidas com o objetivo de predizer o efeito de mutações pontuais sobre a estrutura de uma proteína através do uso de técnicas baseadas na química e física (LAIMER et al., 2015).

Tais ferramentas são agrupadas em duas categorias: as baseadas em sequência e as baseadas em estrutura[1]. A primeira delas utiliza a sequência de aminoácidos das proteínas como dado de entrada (GETOV; PETUKH; ALEXOV, 2016) e embora tais métodos via sequência possam alcançar uma relativa discriminação de mutações de maior impacto, eles não predizem as mudanças estruturais causadas por elas (PETUKH; KUCUKKAL; ALEXOV, 2015). Alternativamente, os métodos baseados em estrutura analisam informações sobre o impacto de mutações em proteínas dentro do seu espaço tridimensional nativo (PIRES; ASCHER; BLUNDELL, 2014a). As abordagens baseadas em sequência e estrutura utilizam diferentes métodos de aprendizado de máquina, como tarefas de classificação (por exemplo, árvores de decisão) (TAN et al., 2006) ou regressão (por exemplo, regressão logística) (ZHAO; RAM, 2008), a fim de predizer as mudanças na energia livre de dobramento dada uma mutação em um aminoácido da proteína.

Embora a maioria dos trabalhos sobre o problema de predição do impacto de mutações pontuais se concentram no desenvolvimento de novos métodos ou heurísticas para resolver as deficiências das abordagens existentes, uma outra técnica, denominada *Ensemble Learning*, visa combinar algoritmos distintos a fim de fornecer predições mais precisas (DIETTERICH, 2000). Um fator relevante a ser mencionado sobre a viabilidade da sua utilização é a dificuldade de seleção de um determinado método *a priori*, tendo em vista que não há como se prever àquele que terá o melhor desempenho para os dados de interesse. Do mesmo modo, o trabalho necessário para teste e comparação de múltiplas abordagens pode tornar o tempo de pesquisa demasiadamente alto para o especialista. Em particular, as abordagens de aprendizado tradicionais tentam construir um único modelo a partir de dados de treinamento. Entretanto, métodos *ensemble* propõem a construção de um conjunto de modelos (ou resultados) e a posterior combinação dos mesmos em uma única saída (ZHOU, 2012). Com isso, a principal proposta desta dissertação é a integração de ferramentas computacionais para a predição do impacto de mutações pontuais em proteínas por meio de uma metodologia *Ensemble Learning*.

1.1 Objetivo geral

Com base na hipótese de que a combinação de resultados individuais melhora o desempenho em tomadas de decisão, esta dissertação tem como objetivo desenvolver uma metodologia cuja finalidade é adotar o conceito de *Ensemble Learning* para combinar em uma única abordagem os resultados de diferentes ferramentas de predição do impacto de mutações pontuais em proteínas.

[1] Neste trabalho adotou-se o termo estrutura quando se refere a uma estrutura tridimensional de proteína.

1.2 Objetivos específicos

Dentre os objetivos específicos podem ser elencados:

- Definir uma metodologia que possibilite o uso de múltiplas técnicas de *Ensemble Learning* a partir de resultados baseados em ferramentas comumente utilizadas;
- Validar a proposta com conjuntos de dados de mutações pontuais definidas experimentalmente;
- Desenvolver uma ferramenta que permita a visualização unificada dos preditores *I-Mutant*, CUPSAT, SDM, mCSM, DUET, iRDP e MAESTRO, juntamente com o método proposto.

1.3 Justificativa

O uso de *Ensemble Learning* é inspirado pelas premissas do *Wisdom of Crowds* (em português, Sabedoria das Massas), a qual refere-se ao fenômeno em que o conhecimento coletivo de uma comunidade é maior do que o de qualquer indivíduo (SUROWIECKI, 2005). Um conceito importante para a análise e avaliação de um método baseado em *ensemble* se deve à diversidade entre os valores envolvidos na combinação e a variabilidade em desempenho, uma das principais motivações para seu uso. A melhoria no desempenho surgiu como uma consequência, mas hoje também pode ser usada como motivação baseada em trabalhos relacionados.

Do ponto de vista biológico, a análise de mutações pontuais desestabilizantes, neutras ou estabilizantes, em conjuntos de proteínas naturais ou modeladas, pode ser extremamente valiosa para refinar a relação entre sequência, estrutura e função de proteínas (MAGLIERY, 2015). De fato, nem todas as mutações são prejudiciais, pelo contrário, algumas são benéficas, pois aumentam a taxa de sobrevivência de uma espécie. Entretanto, se a mutação é prejudicial, pode resultar, por exemplo, em uma doença genética congênita (BETTELHEIM et al., 2012). Dada a sua importância, pesquisas que envolvam melhorias nas estruturas de proteínas podem beneficiar vários segmentos distintos como saúde (desenvolvimento de fármacos, tratamentos, vacinas), indústria em geral (aprimoramento em enzimas digestivas usadas em vários processos), meio-ambiente (alterações em enzimas para a degradação de contaminantes), dentre inúmeros outros (DIAS, 2012).

Considerado-se a relevância desse problema, foram criadas ferramentas computacionais que predizem o impacto de mutações pontuais. Tais modelos implicam no uso de diversas suposições a respeito das probabilidades de substituição de um aminoácido por outro, visando uma aproximação da realidade quando sustentadas por uma melhor acurácia (VERLI et al., 2014). Por isso, quando há diferentes abordagens concorrentes para o problema em questão, um esforço para determinar a mais acurada é inevitável. A

melhor metodologia depende dos dados disponíveis e do conhecimento prévio do especialista (MALINKA, 2015). Dessa forma, através do uso de *Ensemble Learning* busca-se a adoção de abordagens para produzir um resultado final em conjunto potencialmente melhor do que os individuais, levando em consideração que os valores oriundos da tarefa de predição envolvida podem agregar maior generalidade através de consenso.

1.4 Organização do texto

O restante do trabalho está organizado da seguinte maneira:

- o Capítulo 2 consiste em uma fundamentação teórica sobre os conceitos biológicos abordados no decorrer do texto, com um foco acerca de mutações em proteínas. Igualmente, apresenta as principais fontes de dados experimentais de informação biológica que amparam a validação da proposta desta dissertação;
- no Capítulo 3 é feita uma contextualização para o entendimento do restante do trabalho no que se refere ao Aprendizado de Máquina; Nesse mesmo capítulo são apresentados os fundamentos sobre *Ensemble Learning* bem como técnicas de combinação de classificadores;
- o Capítulo 4 compõe a descrição das ferramentas de predição do impacto de mutações pontuais em proteínas utilizadas na abordagem *ensemble* implementada neste trabalho;
- no Capítulo 5 é listado alguns trabalhos já publicados relacionados ao problema desta pesquisa. Estes incluem trabalhos sobre ferramentas de predição que adotam o uso de *Enemble Learning* na sua concepção;
- o Capítulo 6 descreve o método proposto, definido como EN-MUTATE. Da mesma forma apresenta a ferramenta *web* desenvolvida, incluindo uma breve descrição das linguagens de programação e tecnologias adotadas;
- no Capítulo 7, os resultados e respectivas questões de pesquisa propostas na introdução deste trabalho são retomadas e discutidas. A partir das entradas preparadas nos conjuntos de dados baseados em mutações pontuais experimentais, são discutidas as análises da predição sobre esses dados. O EN-MUTATE é então comparado com as demais ferramentas que integram a sua metodologia *ensemble*.
- por fim, no Capítulo 8 são expostas as conclusões desta dissertação. Além disso, são descritas as suas limitações e direções para trabalhos futuros.

2 CONCEITOS BIOLÓGICOS BÁSICOS

Em meados de 1960, com a proposta do Dogma Central da Biologia Molecular, formulado por Francis Crick, tornou-se claro como a matriz genética representada pela sequência nucleotídica poderia determinar o fenótipo, ou seja, as características observáveis de um indivíduo (WATSON et al., 2015). Segundo Lesk e Andrade (2008), um gene corresponde a uma sequência de nucleotídeos ao longo de uma ou mais regiões de uma molécula de DNA (*DeoxyriboNucleic Acid*). Nesse contexto, as regiões codificantes do gene são chamadas de éxons, uma abreviação de "sequências expressadas", de outro modo, os trechos não codificados são chamados de íntrons, uma abreviação para "sequências de intervenção" (NELSON; LEHNINGER; COX, 2008). Dessa forma, o DNA de um gene é então transcrito produzindo uma molécula de RNA (*RiboNucleic Acid*) com sequência complementar. A cada três bases do RNA é traduzido um aminoácido, formando assim uma proteína (VOET; VOET; PRATT, 2014).

A importância dessa transferência de informação genética no entendimento da informação e funções biológicas pode ser exemplificada pelo fato de que ela aborda os três tipos mais comuns de moléculas: o DNA, o RNA e as proteínas (VERLI et al., 2014). Com isso, estabelece um fluxo de informação universal à vida como é conhecida (WATSON et al., 2015). Na Figura 1, as flechas representam a informação transferida quando o DNA direciona a sua própria replicação para produzir novas moléculas, quando o DNA é transcrito em RNA e quando o RNA é traduzido em uma proteína. Esta transferência de informação é conhecida como expressão gênica (VOET; VOET; PRATT, 2014). Nas próximas seções são detalhados os conceitos relacionados a aminoácidos, proteínas e suas mutações, variação de energia livre bem como bancos de dados biológicos.

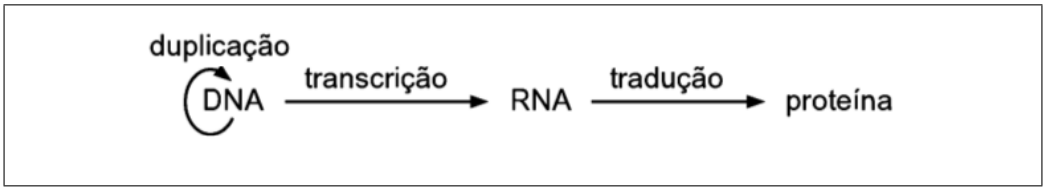


Figura 1: Representação do fluxo de informação em sistemas biológicos.
Fonte: Adaptado de (WATSON et al., 2015).

2.1 Aminoácidos

Os aminoácidos são pequenas moléculas que se caracterizam por terem, pelo menos, um grupo funcional amina (H_2N) e um grupo carboxílico (COOH), um átomo de hidrogênio (H), um carbono alfa ($C\alpha$) e uma cadeia lateral (BARRETT, 2012). Tais moléculas são encontradas nos organismos vivos na forma de peptídeos (2 ou mais aminoácidos) e proteínas (acima de 51 aminoácidos). Nesse cenário, a cadeia lateral é responsável por determinar a identidade do aminoácido, sendo crucial na sua concepção (NELSON; LEHNINGER; COX, 2008). A classificação dos aminoácidos, quanto à cadeia lateral e polaridade, pode ser feita em: neutros, ácidos, básicos, hidrofóbicos, hidrofílicos, polares e apolares (BETTELHEIM et al., 2012). A Figura 2 exibe a estrutura do aminoácido glicina com destaque para a cadeia lateral em vermelho.

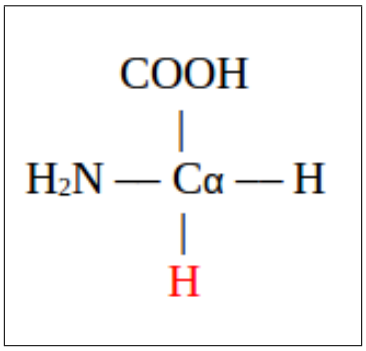


Figura 2: Estrutura do aminoácido glicina.
Fonte: Adaptado de (NELSON; LEHNINGER; COX, 2008).

Com a finalidade de uma padronização geral de uso, a nomenclatura dos aminoácidos é organizada também em abreviações de uma ou três letras. Os nomes dos 20 aminoácidos essenciais que compõem as proteínas, bem como as suas abreviaturas, são descritos na Tabela 1.

Tabela 1: Nomeclatura dos 20 aminoácidos essenciais.

Nome	Abreviatura (1 letra)	Abreviatura (3 letras)
Alanina	A	ALA
Arginina	R	ARG
Asparagina	N	ASN
Aspartato	D	ASP
Cisteína	C	CYS
Glutamina	Q	GLN
Glutamato	E	GLU
Glicina	G	GLY
Histidina	H	HIS
Isoleucina	I	ILE
Leucina	L	LEU

Lisina	K	LYS
Metionina	M	MET
Fenilalanina	F	PHE
Prolina	P	PRO
Serina	S	SER
Treonina	T	THR
Triptofano	W	TRP
Tirosina	Y	TYR
Valina	V	VAL

Fonte: Adaptado de (REID; LOMAS-FRANCIS; OLSSON, 2012).

2.2 Proteínas

As proteínas são polímeros sintetizados pelas células a partir de aminoácidos e constituem o principal produto direto da informação genética a partir da tradução do RNA mensageiro (RNAm) (VERLI et al., 2014). A palavra proteína é derivada do grego *proteios* e significa "de primeira importância" (BETTELHEIM et al., 2012). As proteínas são as moléculas mais abundantes da natureza e praticamente todos os processos da vida dependem dessa classe de moléculas (HARVEY; FERRIER, 2011).

No que se refere a estrutura das proteínas, uma importante característica é a sua conformação especificada pelos valores dos ângulos de torção das duas ligações do esqueleto de cada resíduo bem como os ângulos de torção para toda ligação simples em cada lado da cadeia, sendo fundamental para a função que ela exerce (WATSON et al., 2015). Dessa forma, o termo conformação descreve um arranjo de átomos quimicamente ligados em três dimensões (VOET; VOET; PRATT, 2014). Com isso, embora uma proteína seja uma cadeia linear de aminoácidos covalentes, o seu formato e função são determinados pela estrutura tridimensional estável por ela adotada. Esse formato é determinado pela extensa associação de interações fracas individuais formadas entre aminoácidos que não são necessariamente adjacentes na sequência primária (WATSON et al., 2015). A estrutura de uma proteína é representada, basicamente, de 4 maneiras diferentes (NELSON; LEHNINGER; COX, 2008):

- **Estrutura primária (ou sequência)**: refere-se, em sua essência, à disposição sequencial dos aminoácidos;

- **Estrutura secundária**: compreende as disposições particularmente estáveis dos aminoácidos que dão origem a padrões estruturais recorrentes (alfa-hélices e folhas-beta);

- **Estrutura terciária**: descreve todos os aspectos do dobramento tridimensional de um polipeptídeo;

- **Estrutura quaternária**: existe quando o arranjo espacial de uma proteína tem duas ou mais subdivisões polipeptídicas.

A Figura 3 ilustra um exemplo de cada representação da estrutura da proteína identificada pelo código PDB: 2EJN.

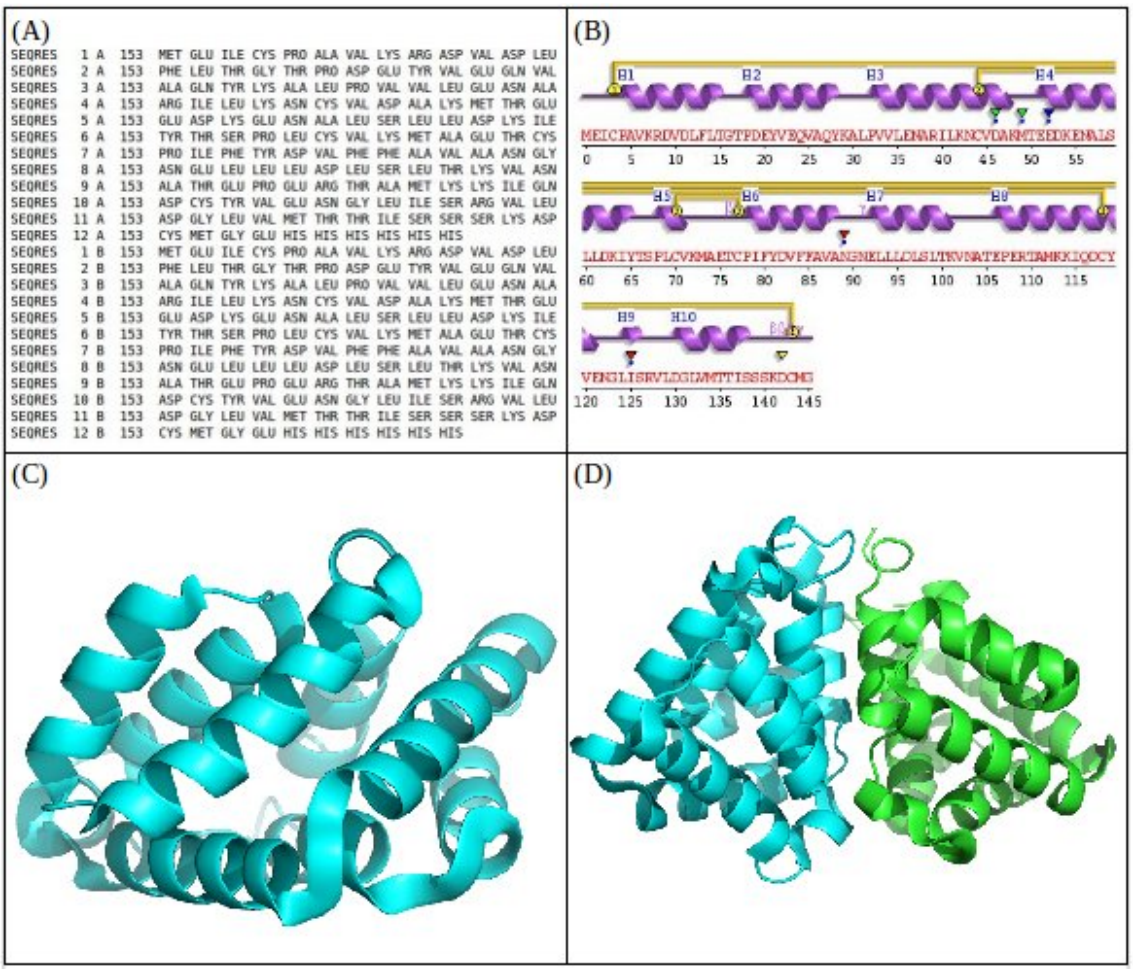


Figura 3: Representações da proteína de código PDB 2EJN: (A) estrutura primária (sequência); (B) estrutura secundária; (C) estrutura terciária; (D) estrutura quaternária.

2.3 Mutações em proteínas

Segundo Horton et. al. (2008), um organismo biológico é um dispositivo de ocorrência natural que se auto-reproduz sendo capaz de manipular matéria, energia e informação. Dessa forma, a vida na Terra é um sistema auto-replicativo complexo, distribuído no tempo e espaço (LESK; ANDRADE, 2008). A teoria da evolução proposta por Charles Darwin (1809-1882) referia-se ao fato de que o sistema estava sujeito a variações, possibilitando o surgimento e a seleção de sistemas variantes mais complexos que se propagam de maneira eficiente (WATSON et al., 2015). Em adição a teoria de Darwin, a descoberta de um elemento ativo na perpetuação da informação genética de todos os organismos

iniciou-se com os experimentos do monge austríaco Gregor Mendel (1822-1884), envolvendo o cruzamento entre ervilhas de diferentes cores. Posteriormente, o trabalho de Thomas Morgan (1866-1945) complementou essa teoria introduzindo a ideia de variação genética por mutações (VOET; VOET; PRATT, 2014).

No estudo da evolução das proteínas, as mudanças que causam alteração permanente de informação genética - como um erro na cópia de uma sequência - são chamadas de mutação (HORTON et al., 2008). A mutação é um processo de mudança genética na estrutura do genoma geralmente causado por um erro na duplicação do DNA, podendo ter consequências deletérias, benéficas ou neutras para o organismo (ALENCAR, 2010). É estimado que, na média, um erro ocorre para cada 10^{10} bases (1 em 10 bilhões) (BETTELHEIM et al., 2012). Na definição de mutação se inclui praticamente todas as alterações permanentes concebíveis na sequência de DNA. As mutações mais simples são as substituições de uma base por outra, ou seja, mutações pontuais (WATSON et al., 2015). Mesmo proteínas muito equivalentes são consideradas importantes pelo fato de que pequenas modificações em sua sequência de aminoácidos podem modificar a sua estrutura e função (DURHAM, 2007). Dessa forma, a fim de explorar características estruturais e funcionais das proteínas, pesquisadores frequentemente introduzem substituições de aminoácidos através de mutagênese dirigida em laboratório (KHAN; VIHINEN, 2010). Tais melhorias podem envolver a mutação de um ou mais aminoácidos, ou seja, as unidades que compõem a proteína, visando um aumento ou diminuição da estabilidade ou flexibilidade com a manutenção ou não da função da proteína (TEILUM; OLSEN; KRAGELUND, 2011). Nesse contexto, o termo estabilidade é definido como a tendência em se manter a conformação nativa de uma proteína. Por exemplo, proteínas nativas são ligeiramente estáveis, pois a variação de energia entre os estados enovelado e desenovelado (ΔG) está na faixa de 4 a 15 Kcal/mol, considerando condições fisiológicas normais (NELSON; LEHNINGER; COX, 2008).

Mutações em determinadas bases do DNA podem alterar o transcrito que será feito a partir do DNA (BETTELHEIM et al., 2012). De acordo com a Figura 1, a primeira etapa da expressão gênica é a transcrição, na qual a região genômica correspondente a um gene é lida pela enzima RNA polimerase, confeccionando uma molécula de RNAm. Nesse processo, as bases pareiam-se: a adenina do DNA se liga à uracila do RNAm, a timina do DNA com a adenina do RNAm, a citosina do DNA com a guanina do RNAm, e assim sucessivamente (ALBERTS et al., 2009). Os aminoácidos são unidos pela ligação peptídica, formando desse modo a proteína, enfileirados na ordem dos códons ao entrar no ribossomo. Uma vez compreendido esse mecanismo, é possível entender que qualquer alteração na sequência do DNA pode levar a alterações na proteína enquanto ela está sendo sintetizada (VOET; VOET; PRATT, 2014). Mutações pontuais em bases do DNA podem provocar alterações na molécula do RNAm, uma vez que é a cópia do DNA, acarretando, desse modo, mudanças nos códons. Esses códons, por sua vez, ao

serem lidos pelo ribossomo terão um aminoácido diferente colocado pelo RNA transportador, potencialmente acarretando em mudanças na proteína, com diversas consequências dependendo da posição e do tipo da mudança (ALBERTS et al., 2009).

Conforme Voet, Voet e Pratt (2014), os tipos de mutações podem ser divididos em:

- **Mutação de inserção/deleção**: alteração genética resultante da adição ou perda de nucleotídeos;
- **Mutação de mudança de quadro**: inserção ou deleção de nucleotídeos no DNA que alteram a sequência de leitura (quadro de leitura) durante a tradução;
- **Mutação não senso (*nonsense*)**: converte um códon que especifica um aminoácido em um códon de parada, provocando o término prematuro da tradução;
- **Mutação pontual**: afeta uma única posição no gene;
- **Mutação silenciosa**: alteram o DNA, mas não mudam os aminoácidos associados;
- **Mutação supressora**: cancela o efeito de outra mutação.

2.4 Variação de energia livre

A formação espontânea de uma ligação entre dois átomos sempre envolve a liberação de uma parte da energia interna dos átomos que não estão ligados e sua conversão à outra forma de energia (WATSON et al., 2015). Quanto mais forte for essa ligação, maior será a quantidade de energia liberada (NELSON; LEHNINGER; COX, 2008). Os fenômenos moleculares da natureza são regidos pela termodinâmica, tanto para as reações químicas na ação da DNA polimerase (enzima que promove a ligação dos nucleotídeos) quanto no enovelamento de proteínas (VERLI et al., 2014). Através disso, as mudanças de energia livre fornecem uma medida da viabilidade energética de uma reação química (HARVEY; FERRIER, 2011). Atualmente, diversos métodos foram desenvolvidos para a obtenção dessas medidas, tais como a perturbação da energia livre, a integração termodinâmica, a energia de interação linear, a metadinâmica e diversas estratégias empíricas voltadas ao pareamento de nucleotídeos ou atracamento molecular (VERLI et al., 2014).

Biologicamente, a maneira mais prática de se expressar a variação de energia livre é por meio do conceito físico-químico em que os átomos se deslocam em direção à sua conformação de equilíbrio, representado pelo símbolo ΔG, uma homenagem ao físico Josiah Gibbs (WATSON et al., 2015). A energia livre de um sistema é representada pela Equação 1, na qual H é a energia de ligação (entalpia) do sistema; T é a sua temperatura em graus Kelvin (K); e S é a entropia, medida de aleatoriedade ou desordem (LODISH et al., 2014).

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

Baseados em conceitos físicos-químicos é possível verificar que a estrutura nativa de uma proteína possui a menor energia livre de qualquer conformação assumida, pois é a conformação mais estável que sua cadeia de aminácidos pode adotar (WATSON et al., 2015). Consequentemente, a relação de ΔG com a direção de uma reação química pode ser resumida em três afirmações (LODISH et al., 2014):

- $\Delta G < \mathbf{0}$: a reação direta tenderá a ocorrer espontaneamente; haverá liberação de energia à medida que a reação acontecer.
- $\Delta G > \mathbf{0}$: a reação direta não ocorrerá espontaneamente; a energia é adicionada ao longo da reação.
- ΔG = **0**: o sistema está em equilíbrio.

Como milhares de calorias estão geralmente envolvidas na quebra de um mol de ligações químicas, a maioria das alterações de energia nas reações químicas são expressadas em quilocalorias por mol (Kcal/mol) (WATSON et al., 2015). A energia livre de dobramento de proteínas é uma característica considerada extremamente importante, sendo diretamente relacionada com a estabilidade da molécula (HORTON et al., 2008). Algumas proteínas são mais estáveis, enquanto outras se desdobram sob uma perturbação muito pequena das suas condições nativas (ZHANG et al., 2012). A maneira mais comum utilizada para se avaliar o impacto de uma mutação pontual é através do $\Delta\Delta G$, representado pela Equação 2. O $\Delta\Delta G$ é definido pela subtração das energias livres resultantes do dobramento da proteína ΔG do tipo mutante (tm) e ΔG do tipo selvagem (ts) (VOET; VOET; PRATT, 2014).

$$\Delta\Delta G = \Delta G_{tm} - \Delta G_{ts} \tag{2}$$

Na construção dos conjuntos de dados desta dissertação, os valores experimentais de $\Delta\Delta G$ foram discretizados em duas e três classes, baseados em Folkman, Stantic e Sattar (2014) bem como Zhao et al. (2014), respectivamente. A classificação binária foi definida como:

- **Mutações desestabilizantes**: $\Delta\Delta G < 0\ Kcal/mol$;
- **Mutações estabilizantes**: $\Delta\Delta G >= 0\ Kcal/mol$.

Com o acréscimo de uma nova classe, a discretização de $\Delta\Delta G$ foi redefinida para uma classificação ternária:

- **Mutações desestabilizantes**: $\Delta\Delta G < -0,5\ Kcal/mol$;

- **Mutações neutras**: $-0,5 \leq \Delta\Delta G \leq 0,5\ Kcal/mol$;
- **Mutações estabilizantes**: $\Delta\Delta G > 0,5\ Kcal/mol$.

No que se refere a métodos computacionais, o uso de funções empíricas têm como objetivo encontrar valores que maximizam a correlação da variação de energia livre ($\Delta\Delta G$) com os dados experimentais de conjuntos que treinam um modelo (chamado conjunto de treinamento) (VERLI et al., 2014). Como resultado, o $\Delta\Delta G$ é o principal valor de saída das ferramentas de predição do impacto de mutações em proteínas encontradas na literatura e abordadas nesta dissertação, sendo empregado para avaliar o impacto na estabilidade de moléculas com a substituição de um aminoácido nativo por outro.

2.5 Bancos de dados biológicos

Bancos de dados especializados, incluindo os de biologia molecular, impõem uma estrutura particular para a informação com o objetivo de separá-las em categorias, sendo originalmente idealizados por grupos de pesquisa individuais (LESK; ANDRADE, 2008). O crescimento da disponibilidade e diversidade de dados de biologia molecular permitiu muitas descobertas e avanços em diferentes campos relacionados a biologia de sistemas (WANG et al., 2005). Embora as bases de dados confiáveis que disponibilizam informação estrutural e funcional de proteínas tenham um considerável crescimento ao longo de cada ano[1], não havia nenhuma base de dados *online* acessível a toda a comunidade acadêmica antes dos anos 2000. No entanto, atualmente é possível utilizar, inclusive, o resultado de uma pesquisa a uma base de dados biológica como entrada para outro programa, uma vez que grande parte das ferramentas de bioinformática oferecem recursos para iniciar processos como esse, por exemplo, a obtenção de uma estrutura de proteína através do seu código de identificação (LESK; ANDRADE, 2008).

2.5.1 *Protein Data Bank*

O *Protein Data Bank* (PDB) é um repositório público com dados de macromoléculas biológicas gerenciado pelo RCSB (*Research Collaboratory for Structural Bioinformatics*) (BERMAN et al., 2000). Sua base de dados contém, em março de 2017, aproximadamente, 125 mil estruturas armazenadas. O conteúdo de um arquivo PDB é descrito em um formato padrão aceito por grande parte das ferramentas de visualização de proteínas, agregando informações espaciais e sequenciais para cada um dos átomos (separados por resíduo) de uma dada proteína.

As Figuras 4 e 5 exibem trechos do arquivo de estrutura da proteína de código PDB: 1A98. Os conjuntos de linhas não exibidos foram substituídos por "...".

[1] PDB. Yearly Growth of Total Structures. Gráfico. Disponível em: <http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total>. Acessado em: 10/03/2017.

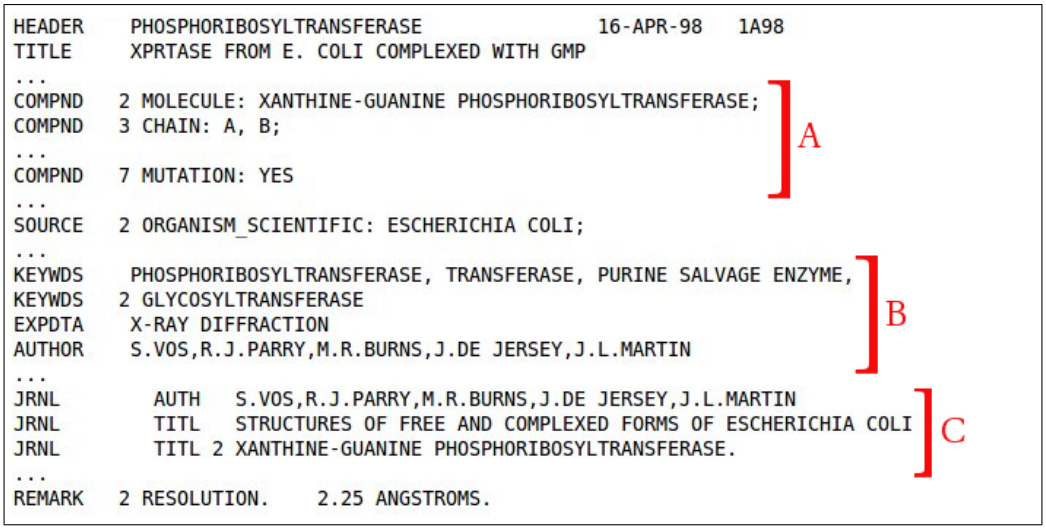


Figura 4: Trechos de um arquivo PDB padrão: cabeçalho.

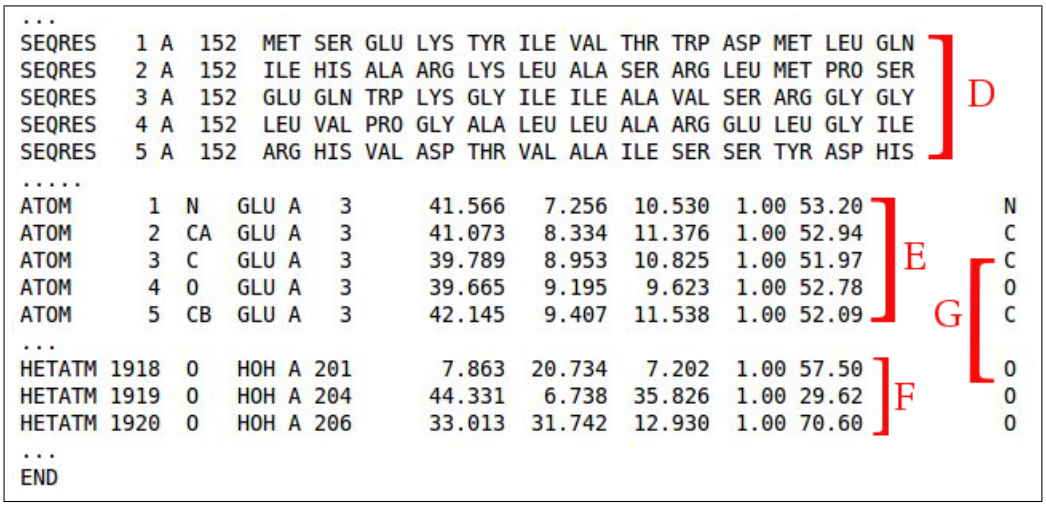


Figura 5: Trechos de um arquivo PDB padrão: estrutura primária e terciária.

A Figura 4 descreve parte do cabeçalho do arquivo PDB. Inicialmente, é descrito o grupo referente à estrutura (HEADER) e a sua identificação (TITLE). Em "A" são exibidos os registros (COMPND) contendo detalhes da molécula, suas cadeias bem como um indicativo, nesse caso, de que se trata de uma proteína do tipo mutante. Abaixo de "A", o termo SOURCE especifica a origem da molécula. No bloco "B" são exibidas as palavras-chave (KEYWDS) de indexação, detalhes sobre a técnica utilizada (EXPDTA) e os autores do experimento (AUTHOR). Destacado em "C" podem ser vistos detalhes sobre a publicação gerada a partir dessa estrutura (JRNL). Nos comentários (REMARK) é mostrado, em *Angstroms*, a resolução mais alta utilizada na criação do modelo.

Posteriormente, a Figura 5 apresenta em "D", 5 das 12 linhas que definiram nesse exemplo a sequência de aminoácidos da estrutura primária da proteína (SEQRES). Iden-

tificado por "E", consta o bloco de coordenadas atômicas (ATOM) referente ao aminoácido de posição 3, glutamina (GLU), na qual é possível observar a identificação dos seus átomos da cadeia principal (N, CA, C, O, CB), cadeia pertencente (A), bem como as suas coordenadas (x, y, z), ocupação e fator-B (fator de temperatura para cada átomo). No bloco destacado em "F" estão representados os heteroátomos (HETATM), que constituem as coordenadas de átomos dentro de grupos "não-padrão". Esses registros são usados para as moléculas de água e os átomos representados nos grupos HET. Finalmente, em "G", são exibidos os símbolos dos elementos (representação em uma letra).

2.5.2 *ProTherm*

ProTherm é uma coleção pública de dados de parâmetros termodinâmicos de proteínas. Neles se incluem mudanças de energia livre, variação de entalpia, mudanças na capacidade de calor bem como transição de temperatura, tanto para proteínas do tipo selvagem quanto para proteínas do tipo mutante (BAVA et al., 2004). Sua base de dados também contém informações sobre a estrutura secundária das proteínas e acessibilidade dos resíduos, assim como medições e o método utilizado em cada experimento. O seu uso é citado em grande parte dos artigos relacionados a ferramentas de predição de estabilidade em mutações pontuais. As informações contidas no *ProTherm* servem para a compreensão da estrutura e estabilidade das proteínas, servindo como elemento validador de várias abordagens de predição do impacto de mutações em proteínas. As Figuras 6 e 7 mostram um exemplo de pesquisa de dados termodinâmicos através do *ProTherm*.

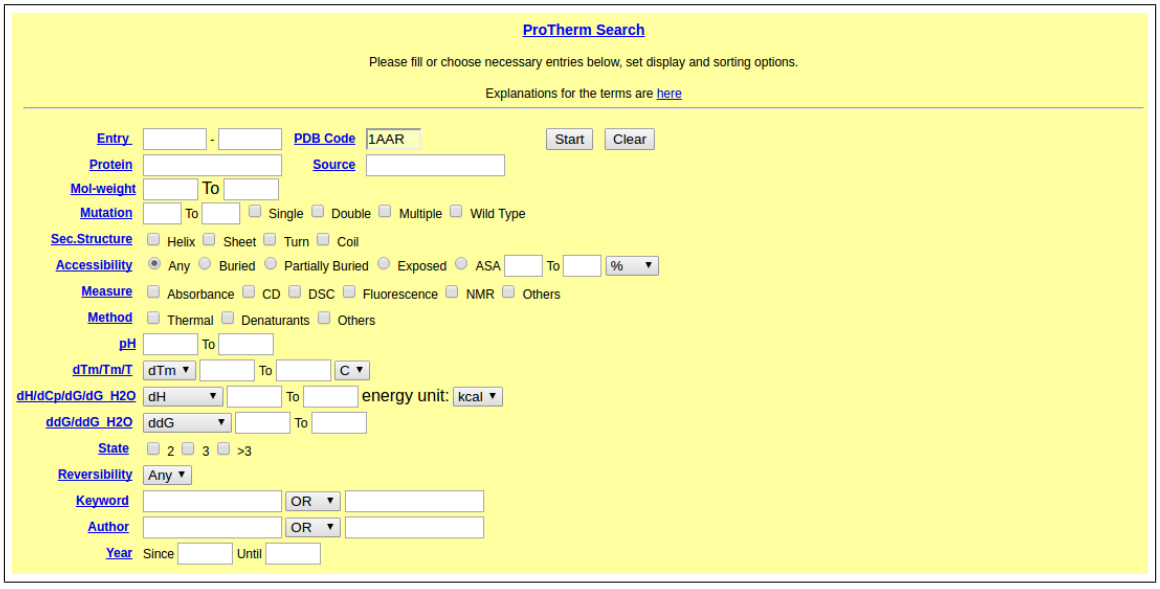


Figura 6: Exemplo de uso: pesquisa via *ProTherm*.
Fonte: http://www.abren.net/protherm/protherm_search.php

Dentre as opções de filtro de pesquisa exibidas na Figura 6, merecem destaque as entradas rotuladas por: "*PDB Code*", "*Mutation*", "*Measure*" e "*Method*". Em "*PDB Code*", é possível pesquisar dados oriundos de uma proteína específica através do seu código

PDB, como exemplo, a proteína identificada por 1AAR. O próximo campo, "*Mutation*", permite a inserção de um intervalo de mutações relacionado às posições dos resíduos na sequência da proteína. Também é permitido a escolha por somente mutações pontuais únicas ou duplas. Por fim, os campos "*Measure*" e "*Method*" referem-se, respectivamente, aos experimentos realizados para medir os parâmetros termodinâmicos (espectroscopia de fluorescência, dicroísmo circular, calorimetria diferencial de varredura, absorvância, RMN, etc) e o método experimental de desnaturação (palavras-chave: térmica, ureia, etc).

Search Condition
PDB: 1AAR
Sorting by ddg,

Entry	Protein	Source	Mutation	ddG	pH	Measure	Method	REFERENCE
5984	Ubiquitin	Bovine	R 42 E	1.63	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5986	Ubiquitin	Bovine	H 68 E	0.77	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5985	Ubiquitin	Bovine	H 68 Q	0.55	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5979	Ubiquitin	Bovine	K 6 E	0.53	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
2418	Ubiquitin	Bovine	F 45 W	0.32	5.00	CD	Thermal	BIOCHEMISTRY 32, 7054-7063 (1993) PMID: 8392867
5980	Ubiquitin	Bovine	K 6 Q	0.26	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5987	Ubiquitin	Bovine	R 72 Q	-0.33	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5983	Ubiquitin	Bovine	K 29 N	-1.48	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5982	Ubiquitin	Bovine	K 29 Q	-1.67	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5981	Ubiquitin	Bovine	K 27 Q	-1.91	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
5978	Ubiquitin	Bovine	WILD	NULL	5.00	CD	Thermal	BIOCHEMISTRY 38, 16419-16423 (1999) PMID: 10600102
11623	Ubiquitin	Bovine	WILD	NULL	4.00	CD	Thermal	BIOCHEMISTRY 40 10317-10325 (2001) PMID: 11513610
11622	Ubiquitin	Bovine	WILD	NULL	4.00	CD	Thermal	BIOCHEMISTRY 40 10317-10325 (2001) PMID: 11513610

Figura 7: Exemplo de uso: resultados da pesquisa via *ProTherm*.
Fonte: http://www.abren.net/protherm/protherm_result.php

O resultado da pesquisa (Figura 7) é desse modo exibido após a submissão do formulário. A partir dos dados de entrada supra citados, a base de dados traz informações sobre a identificação da proteína, como, o código *ProTherm*, o nome, a origem e a publicação de referência, sendo que a correspondência é feita através de entradas dos bancos de dados biológicos: PDB, *UniProt*, *Swiss-Prot* e do próprio *ProTherm*.

A última atualização do *ProTherm* ocorreu em fevereiro de 2013. Atualmente, março de 2017, a sua base de dados possui um total de 25820 entradas cadastradas para 740 proteínas distintas. Entretanto, apenas aproximadamente 30% dos resultados são valores de $\Delta\Delta G$, principal métrica na avaliação do impacto de mutações pontuais. Os conjuntos de dados abordados nos experimentos desta dissertação foram composto por mutações pontuais não redundantes e ordenadas de maneira decrescente (mais atuais) com base no campo "*Entry*" da pesquisa do *ProTherm*. Foram consideradas inutilizáveis as entradas de mutações em que a alteração entre a estrutura nativa e a mutante se dava, por exemplo, pelo valor de pH ou temperatura, e não por substituições de aminoácidos.

3 APRENDIZADO DE MÁQUINA: UMA ABORDAGEM *ENSEMBLE*

O Aprendizado de Máquina (AM) é uma área da Inteligência Artificial (IA) que compreende o desenvolvimento de técnicas computacionais capazes de adquirir conhecimento de forma automática com base em experiências de problemas anteriores (WEISS; KULIKOWSKI, 1991). Atualmente, as técnicas de AM estão organizadas em três focos de pesquisa diferentes: estudos orientados à tarefas, simulação cognitiva e análise teórica (MICHALSKI; CARBONELL; MITCHELL, 2013). Com isso, os algoritmos de AM têm sido comumente utilizados em tarefas bastante distintas, podendo ser organizadas de acordo com diferentes critérios. O principal deles diz respeito ao paradigma de aprendizado a ser adotado para lidar com a tarefa preditiva (FACELI et al., 2011). No entendimento de Han, Pei e Kamber (2006), com base nesse critério, as tarefas de aprendizado podem ser divididas em:

- **Aprendizado supervisionado (preditivo)**: é aquele que utiliza dados com uma classe especificada. Geralmente são técnicas que tentam predizer a qual classe uma instância desconhecida pertence baseado nos exemplos utilizados em seu treinamento. Existem dois tipos de modelos de predição: classificação e regressão;

- **Aprendizado não supervisionado (descritivo)**: é aquele que utiliza instâncias sem a determinação de um atributo classe. É utilizado frequentemente para a busca de padrões entre dados. As técnicas descritivas são: associação, agrupamento e sumarização.

Segundo Faceli et al. (2011), o problema de AM pode ser formulado como um problema de procura em um espaço de possíveis soluções. A análise de dados oriundos de AM não só reduz a carga de trabalho para o experimentalista, mas, também, busca objetividade e consistência em grandes conjuntos de dados (DANUSER, 2011). Em uma variedade de campos de atuação, os dados são coletados e acumulados em um ritmo acelerado, tornando necessária a aplicação de teorias computacionais e ferramentas para ajudar os especialistas a extrair informação útil de dados brutos (TAN et al., 2006).

A Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases*, KDD) visa um processo geral de conversão de dados brutos em informações úteis na qual a mineração de dados (em inglês, *Data Mining*) é uma parte integral do seu processo (FAYYAD et al., 1996). De acordo com Tan et al. (2006), a mineração de dados geralmente decorre em duas fases: na primeira delas, denominada fase de treinamento, uma coleção de amostras de dados é usada para aprender com estruturas e relações inerentes a esses dados; na segunda fase, chamada de teste, o modelo definido é então aplicado em novas amostras de dados para predizer certas propriedades. Em geral, esses padrões são extraídos de relacionamentos implícitos entre os dados a serem analisados. À vista disso, os padrões encontrados devem gerar um conhecimento claro e utilizável para o apoio às decisões (HAN; PEI; KAMBER, 2006). Assim, o objetivo principal de qualquer técnica de mineração de dados é a generalização a partir de alguns exemplos de treinamento, a fim de realizar predições sobre grandes conjuntos de amostras de dados que não foram observadas durante o treinamento (RIDDER; RIDDER; REINDERS, 2013).

3.1 Classificação de dados

Para Han, Pei e Kamber (2006), classificação é o processo de encontrar um modelo (ou função) que descreve e distingue classes de dados ou conceitos. Nessas abordagens são gerados conjuntos de treinamento com exemplos representativos de acordo com as classes predefinidas. Com isso, o algoritmo de AM infere automaticamente as regras para discriminar as classes, podendo ser aplicadas para o conjunto de dados completo (BISHOP, 2006). Dessa maneira, cada instância que fornece a entrada para o AM é caracterizada por seus valores em um conjunto fixo e pré-definido de atributos (HASTIE; TIBSHIRANI; FRIEDMAN, 2011). Nesses atributos são incluídos os denominados numéricos (também chamados de atributos contínuos), sendo medidas de valores reais ou inteiros, bem como, atributos ditos nominais (às vezes chamados de categóricos), os quais assumem valores em um conjunto pré-determinado e finito de possibilidades (ZAKI; MEIRA, 2014).

As tarefas de classificação devem ser precedidas por uma análise de relevância na qual é possível identificar atributos que são significativamente relevantes para o seu processo de aprendizado (HAN; PEI; KAMBER, 2006). Por fim, os modelos de classificação são usados para prever a classe de objetos para os quais o rótulo da classe é desconhecido (ROKACH; MAIMON, 2014).

Em Bioinformática, a classificação é uma das abordagens mais populares no entendimento da relação entre características de vários objetos, uma vez que trata de um processo que encontra propriedades comuns entre um conjunto de dados e os organiza em diferentes classes (WANG et al., 2005). Nesta dissertação foram executados experimentos de mineração de dados com técnicas de classificação visando permitir um melhor entendimento sobre o impacto que uma mutação pontual pode causar na estrutura de uma pro-

teína, dessa forma, reduzindo o espaço de possibilidades a ser considerado na indução de mutações pontuais pelo especialista em laboratório.

3.2 Métricas de avaliação

Alguns desafios são geralmente encontrados em tarefas de classificação, por exemplo, classificadores que sofrem com distribuições de classe desequilibradas. Esse tipo de problema ocorre quando o número de exemplos que representam uma classe é muito menor do que os das outras classes (WANG; YAO, 2013). Considerando um conjunto de dados cuja razão de desequilíbrio é 1:100, para cada exemplo da classe positiva há 100 exemplos de classe negativa. Dessa forma, um classificador que tenta maximizar a acurácia da sua regra de classificação poderá obter uma acurácia de 99% apenas pela desconsideração dos exemplos positivos e com a classificação de todas as instâncias como negativos (GALAR et al., 2012). O estado da arte em métodos de AM são otimizados para aprender com o menor número de instâncias de treinamento sem perder a versatilidade na sua aplicação (SOMMER; GERLICH, 2013). A Tabela 2 descreve as principais métricas de avaliação da predição por um classificador.

Tabela 2: Métricas de avaliação de um classificador.

Métrica	Fórmula
Precisão	$\frac{VP}{VP+FP}$
Revocação	$\frac{VP}{VP+FN}$
Medida-F	$\frac{2\ X\ precisão\ X\ revocacão}{precisão + revocacão}$
Acurácia	$\frac{VP+VN}{VP+VN+FP+FN}$

Fonte: Adaptado de (TAN et al., 2006)

Conforme Tan et. al (2006), as propriedades das métricas de avaliação são descritas como:

- **Verdadeiro positivo (VP)**: número de instâncias verdadeiras positivas classificadas corretamente;
- **Verdadeiro negativo (VN)**: número de instâncias negativas classificadas corretamente;
- **Falso positivo (FP)**: número de instâncias negativas classificadas erroneamente como positivas;
- **Falso negativo (FN)**: número de instâncias positivas classificadas erroneamente como negativas.

A Precisão (do inglês, *Precision*), trata da proporção de exemplos positivos classificados corretamente entre todos preditos como positivos. Semelhantemente, a Revocação (também chamada em inglês de *Recall*) corresponde à taxa de acerto da classe positiva (FACELI et al., 2011). A Medida-F considera os valores de Precisão e Revocação calculados e a Acurácia (acerto total) calcula a taxa de instâncias que foram classificadas corretamente durante o processo de validação. Em ambas as métricas, os maiores valores indicam os melhores modelos (WITTEN; FRANK; HALL, 2011). Os termos apresentados nas fórmulas da Tabela 2 podem ser resumidos através de uma Matriz de Confusão. Essa representação, ilustrada na Figura 8, é uma opção útil para analisar o quão bem um classificador pode reconhecer tuplas de diferentes classes, uma vez que VP e VN representam as predições realizadas corretamente, enquanto FP e FN os erros de predição.

		Classe predita	
		Sim	Não
Classe atual	Sim	VP	FN
	Não	FP	VN

Figura 8: Matriz de confusão. A diagonal principal representa os acertos.
Fonte: Adaptado de (HAN; PEI; KAMBER, 2006).

3.3 *Ensemble Learning*

Segundo Dietterich (2000), métodos *Ensemble* são algoritmos de aprendizado que constroem um conjunto de classificadores e, posteriormente, classificam novos dados tendo um voto de suas previsões. A principal característica desses métodos não é apenas melhorar o desempenho geral de predição, mas, também, a capacidade de generalização (HO, 2002). Em particular, os métodos de aprendizado *ensemble* consistem em gerar um conjunto de soluções candidatas diversas, seja por otimização baseada em diferentes dados ou até mesmo em algoritmos distintos (KUNCHEVA, 2004).

Abordagens de *Ensemble Learning* têm sido aplicadas em questões de grande importância que envolvem áreas como a economia, a logística e até mesmo a medicina (POLIKAR, 2012). Um dos primeiros trabalhos encontrados na literatura sobre *Ensemble Learning* é o de Dasarathy e Sheela (1979), nele são exploradas as possibilidades de atingir um melhor desempenho de um sistema de reconhecimento por meio da implantação de uma abordagem de classificação constituída por dois ou mais componentes classificadores que pertencem a diferentes categorias. Semelhantemente, o artigo de Hansen e Salamon (1990) aplica uma série de procedimentos para a análise e melhoria de tarefas de classificação através de redes neurais. A ideia básica é classificar um determinado padrão de entrada por meio da predição de cada cópia da rede para posteriormente utilizar um

esquema de consenso com o propósito de decidir a classificação coletiva através de voto. O desempenho do classificador atingido pelo consenso majoritário foi baseado na taxa de erro médio de cada rede.

Os modelos de aprendizado oriundos do *ensemble* combinam suas decisões, seus algoritmos ou até mesmo dados distintos, a fim de se atingir predições mais acuradas. Em um modelo de AM baseado em *ensemble*, busca-se a melhor predição comparada a qualquer modelo único utilizado separadamente (SOARES, 2015). A Figura 9 ilustra uma arquitetura básica abordada pelo *ensemble*.

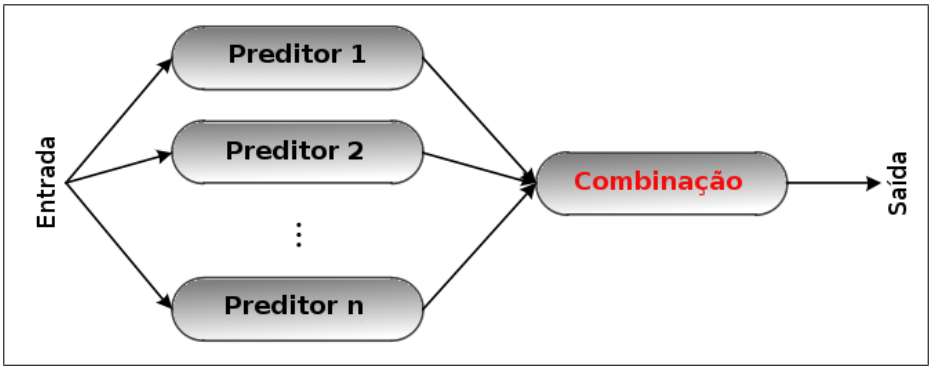


Figura 9: Arquitetura tradicional de *Ensemble Learning*.
Fonte: Adaptado de (ZHOU, 2012).

Da mesma forma, o uso da combinação de resultados ao invés de modelos individuais para a tomada de decisão consensual está embasado nos pressupostos da estatística, computabilidade e representabilidade, ilustrados na Figura 10. Com isso, é possível inferir um veredito mais coerente (DIETTERICH, 2000):

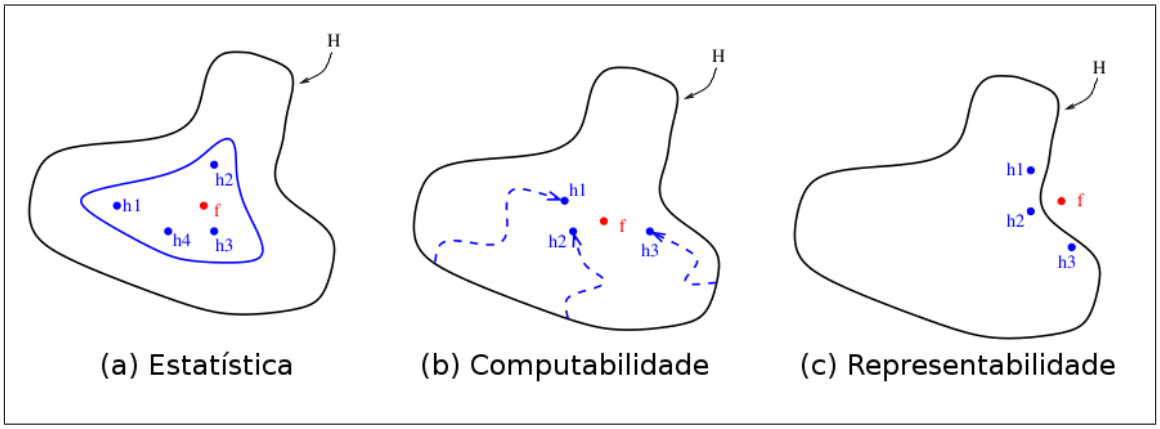


Figura 10: Três razões fundamentais pelas quais se justifica o uso de *ensemble*, sendo f a solução ótima em um espaço H de possibilidades *{h1, h2, h3 ..., hn}*.
Fonte: Adaptado de (DIETTERICH, 2000).

- **Estatística**: pode ser considerado como a procura de um espaço H de hipóteses para identificar a melhor hipótese no espaço (*h1*, *h2*, *h3*, *h4*), conforme destaca a

Figura 10 (a). No entanto, sem dados suficientes, o algoritmo de aprendizado pode encontrar muitas hipóteses diferentes em *H* para os dados de treinamento.

- **Computabilidade**: vários algoritmos de aprendizado operam através da realização de alguma forma de busca local, representação da Figura 10 (b). A saída desses algoritmos geralmente fornece uma solução aproximada para o problema que pode variar entre várias execuções dada a natureza estocástica dos métodos de busca local. Um *ensemble* construído para executar uma busca local a partir de diferentes pontos de partida pode fornecer uma melhor aproximação à verdadeira função desconhecida, isso comparado a qualquer um dos classificadores individuais.

- **Representabilidade**: na maioria das aplicações de AM, a função verdadeira *f* não pode ser representada por nenhuma das hipóteses em *H*, ilustração da Figura 10 (c). Por exemplo, ao formar somas ponderadas de hipóteses tiradas de *H*, pode ser possível expandir o espaço de funções representáveis. No entanto, com uma amostra de treinamento finita, esses algoritmos explorarão apenas um conjunto limitado de hipóteses e a pesquisa será interrompida caso uma hipótese que se ajuste aos dados de treinamento for encontrada.

3.4 Combinação de classificadores

Segundo Han et al. (2006), a combinação de classificadores consiste em uma série de k modelos de aprendizado (ou classificadores de base), M_1, M_2, ... , M_k, com o objetivo de criar um modelo de classificação composto melhorado. Um determinado conjunto de dados, D, é utilizado para gerar k, D_1, D_2,. . . , D_k conjuntos de formação, em que D_i (1 $\leq i \leq k$ - 1) é usado para gerar um classificador M_i. Dada uma nova tupla de dados para classificar, cada um dos os classificadores base votam devolvendo a predição da classe. A combinação de classificadores deve ser feita com critério, de modo a diminuir a sua redundância uma vez que que modelos redundantes aumentam a sua complexidade e o tempo de processamento (BISHOP, 2006). Desse modo, identificar qual a melhor combinação de classificadores pode ser considerado um problema combinatorial (HODGE; AUSTIN, 2004).

Em tais casos, combinar as saídas de vários preditores pode reduzir o risco da escolha de um resultado ruim, em razão de que o resultado dessa combinação pode, ou não, superar o desempenho do melhor classificador do conjunto, reduzindo o risco geral de se tomar uma decisão menos acurada (POLIKAR, 2006). Nesse contexto, os classificadores que implementam algoritmos potencialmente diferentes oferecem informação complementar sobre os padrões a serem classificados (DIETTERICH, 2000). Portanto, a combinação das predições realizadas por esses algoritmos podem aumentar a qualidade do processo de classificação. Na implementação de *ensemble learning*, diferentes técnicas foram pro-

postas a fim de se combinar um grupo de hipóteses em uma única decisão (KUNCHEVA, 2004). Neste trabalho são abordados: *Bagging* (BREIMAN, 1996), *Boosting* (FREUND; SCHAPIRE et al., 1996), *Stacking* (DZEROSKI; ZENKO, 2004), *Cascading* (GAMA; BRAZDIL, 2000) e *Voting* (LIN et al., 2003; KITTLER et al., 1998).

3.4.1 *Bagging*

Segundo Breiman (1996), *Bagging* é uma maneira relativamente fácil de melhorar um método existente, pois tudo o que precisa ser feito é adicionar um laço (*loop*) que seleciona uma determinada amostra e a envia para um procedimento em uma extremidade que faz a agregação. O *Bagging* é uma popular técnica que consiste em ter cada modelo na votação *ensemble* com pesos iguais (GALAR et al., 2012). Com isso, não é necessário adaptar a fórmula de atualização dos pesos nem alterar os cálculos do algoritmo (BREIMAN, 1996).

A Figura 11 ilustra o modelo conceitual de funcionamento do *Bagging*.

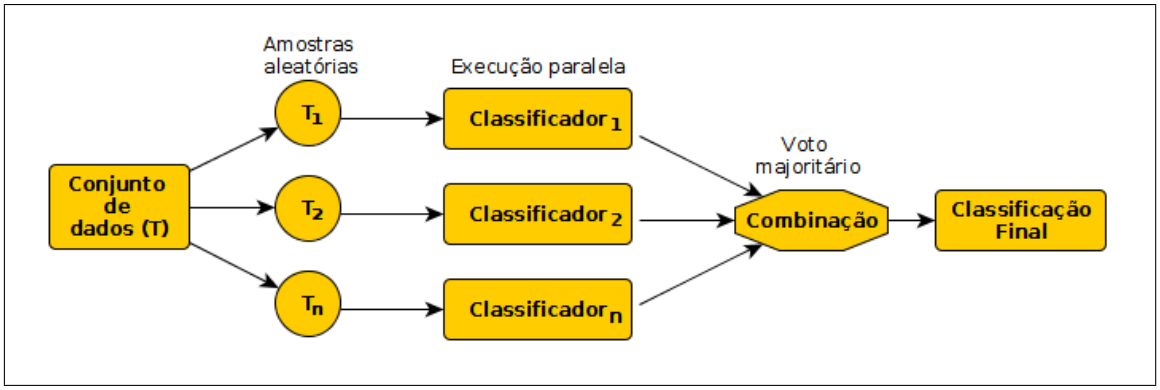


Figura 11: Modelo conceitual: *Bagging*.
Fonte: Adaptado de (SHAFIEI; JAZAYERI-RAD, 2012).

A fim de promover a variância dos modelos, o *Bagging* treina cada um deles usando um subconjunto retirado aleatoriamente do conjunto de dados de treinamento (ZHANG et al., 2016). A ideia do conjunto de teste é que ele seja formado por amostragens independentes com a mesma distribuição subjacente que deu origem ao conjunto de aprendizado (HAN; PEI; KAMBER, 2006). À vista disso, muitas abordagens baseadas em *Bagging* têm sido desenvolvidas para lidar com problemas de desequilíbrio de classe devido à sua simplicidade e capacidade de generalização.

3.4.2 *Boosting*

Boosting é uma técnica que envolve a implementação incremental de um *ensemble* através da formação de cada novo modelo enfatizando as instâncias mal classificadas pelos modelos anteriores (FREUND; SCHAPIRE et al., 1996). Em abordagens baseadas nessa técnica também são incluídos algoritmos que incorporam a prática de pré-processamento de dados. Desse modo, tais métodos podem alterar a distribuição dos

pesos utilizados para treinar um classificador a cada nova iteração (WITTEN; FRANK; HALL, 2011).

Tanto o *Bagging* quanto o *Boosting* adotam abordagens bastante parecidas, entretanto derivam os modelos individuais de maneiras diferentes. No *Bagging*, os modelos recebem o mesmo peso, enquanto no *Boosting* a ponderação é usada para dar influência aos modelos mais bem sucedidos, dependendo do sucesso de suas predições anteriores. À vista disso, os modelos são treinados em sequência (WITTEN; FRANK; HALL, 2011).

O esquema de funcionamento do *Boosting* é representado pela Figura 12.

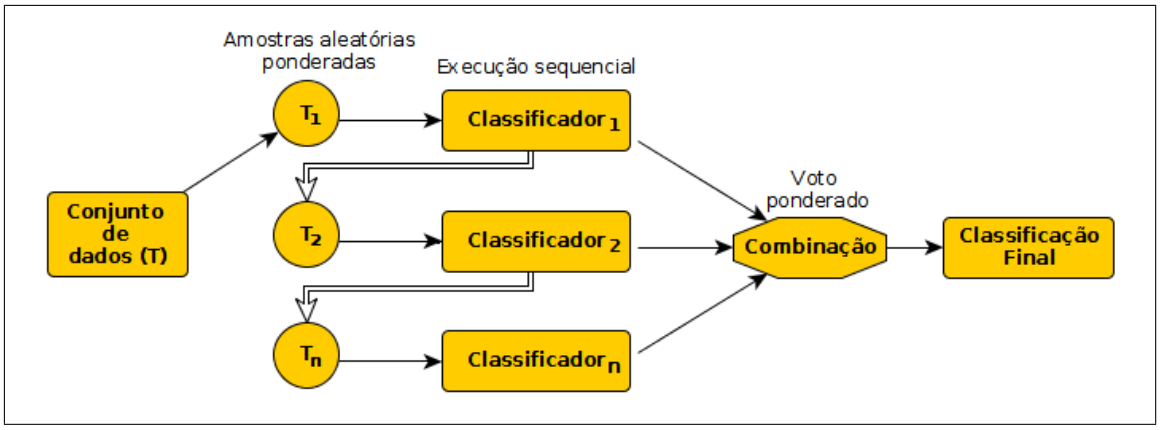


Figura 12: Modelo conceitual: *Boosting*.
Fonte: Adaptado de (SHAFIEI; JAZAYERI-RAD, 2012).

3.4.3 *Stacking*

Stacking (do português, empilhamento), consiste em múltiplos classificadores gerados a partir de diferentes algoritmos de aprendizado aplicados sobre um único conjunto de dados, os quais são tipicamente combinados por voto majoritário ou ponderado (DZEROSKI; ZENKO, 2004). Nesse cenário, os classificadores individuais que compõem o *stacking* são chamados de primeiro nível, enquanto o combinador é chamado de classificador de segundo nível ou meta classificador (ZHOU, 2012). A ideia básica é treinar os modelos do primeiro nível utilizando o conjunto de dados de treinamento original e, em seguida, gerar um novo conjunto de dados para treinar o modelo do segundo nível, nas quais o conhecimento adquirido pelos classificadores de primeiro nível são considerados como recursos de entrada, sendo finalmente utilizado para inferir a classe das instâncias do conjunto de teste (WITTEN; FRANK; HALL, 2011).

Ao contrário do *Bagging* e do *Boosting*, por exemplo, o *Stacking* normalmente não é utilizado para combinar modelos do mesmo tipo, como um conjunto de árvores de decisão. Em vez disso, ele é aplicado a modelos construídos por diferentes algoritmos de aprendizado (WITTEN; FRANK; HALL, 2011). A Figura 13 ilustra o modelo conceitual de funcionamento do *Stacking*.

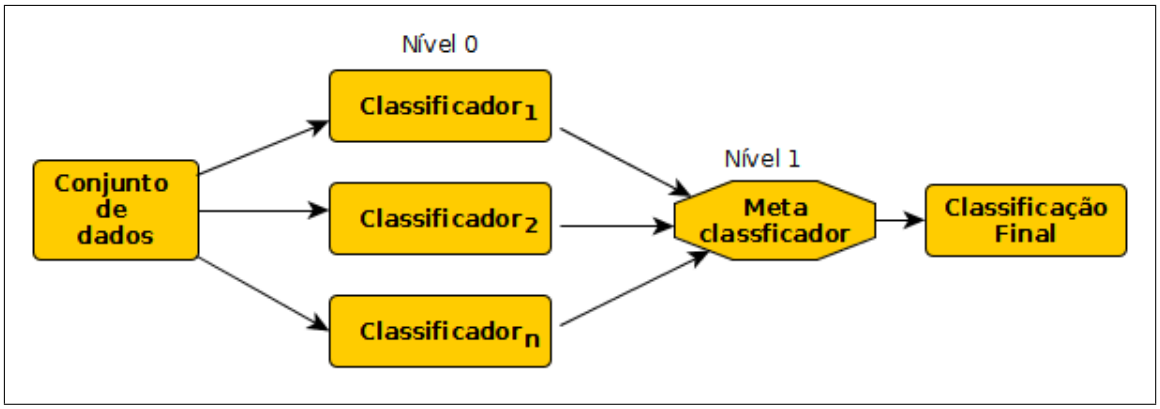


Figura 13: Modelo conceitual: *Stacking*.
Fonte: Adaptado de (EBRAHIMPOUR et al., 2010).

3.4.4 *Cascading*

O *Cascading* está na categoria de sistemas de múltiplos estágios. Seu viés indutivo é que o conceito pode ser explicado por um pequeno número de regras com um conjunto adicional de exceções (GAMA; BRAZDIL, 2000). Na primeira etapa, é construída uma regra simples para todo o conjunto de treinamento usando um classificador generalizador (OLIVEIRA; BRITTO; SABOURIN, 2005). Dado que esses classificadores podem cometer erros em diferentes partes do espaço de entrada, eles se complementam em um esquema *ensemble* capaz de superar os classificadores individuais (KAYNAK; ALPAYDIN, 2000). A Figura 14 ilustra o seu modelo conceitual de funcionamento do *Cascading*.

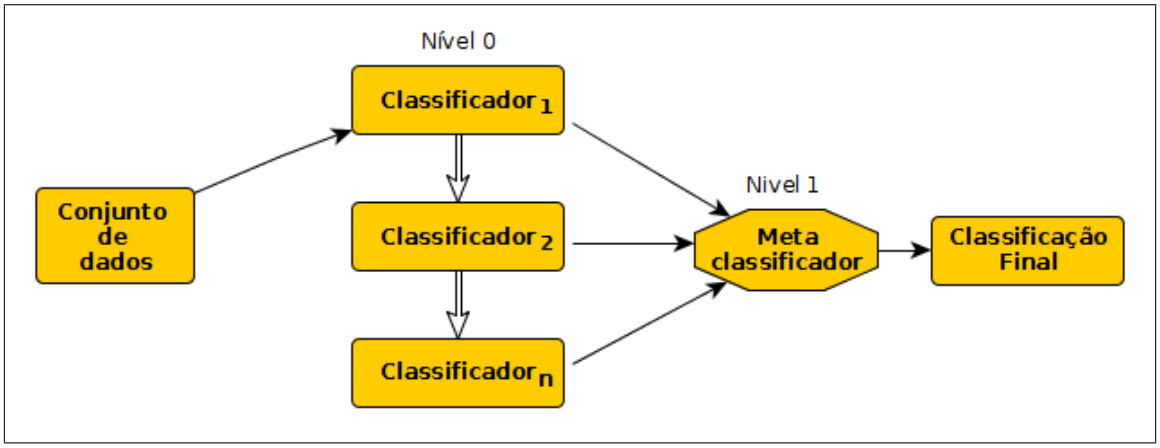


Figura 14: Modelo conceitual: *Cascading*.
Fonte: Adaptado de (KARLOS et al., 2016).

A técnica *Cascading* segue a teoria da generalização empilhada (*stacked generalization*). Em cada nível, um ou mais classificadores escolhidos examinam um conjunto de dados independentemente e realizam a sua predição (GAMA; BRAZDIL, 2000). Essas predições são adicionadas ao conjunto de dados inicial e passadas para o próximo nível. Como resultado, o classificador do último nível é fornecido com um conjunto de dados

ampliado e sua saída é claramente orientada e afetada pelas decisões dos níveis inferiores (KARLOS et al., 2016). Em algumas técnicas de *Cascading*, há uma sequência de classificadores ordenados em termos crescentes de complexidade e especificidade, uma vez que os primeiros classificadores tendem a ser mais simples e gerais, enquanto que os posteriores mais complexos e específicos (KAYNAK; ALPAYDIN, 2000).

3.4.5 *Voting*

Conforme Kuncheva (2004), existem basicamente três principais versões de votação em que cada votante define um voto para a decisão de uma dada entrada de acordo com a sua decisão individual:

- ***Majority voting***: exige pelo menos mais da metade (50% + 1) do número de votantes para uma determinada classe;

- ***Plurality voting***: a decisão mais popular, ou seja, aquela com a maioria dos votos, é então escolhida;

- ***Weighted voting***: é semelhante à abordagem anterior, entretanto, torna-se possível atribuir pesos para cada votante.

3.5 Comparativo entre os classificadores

A Tabela 3 resume as principais características apresentadas por cada classificador *ensemble* abordado neste capítulo.

Tabela 3: Comparativo entre os classificadores *ensemble*: principais características.

Classificador	Principais características	Referência
Bagging	Define amostras separadas do conjunto de dados de treinamento bem como um classificador para cada amostra. Os resultados desses classificadores são então combinados (votação majoritária).	(BREIMAN, 1996)
Boosting	Começa com um classificador de base que é preparado com os dados de treinamento, logo um segundo classificador é então criado com o foco nas instâncias que o primeiro classificador errou. O processo continua a adicionar classificadores até que um limite seja atingido no número de modelos ou valor de acurácia.	(FREUND; SCHAPIRE et al., 1996)

Stacking	Múltiplos algoritmos distintos (nível 0) são preparados com os dados de treinamento e posteriormente um meta-classificador combinador (nível 1) é treinado para realizar uma predição final utilizando como entrada todas as predições dos outros algoritmos	(DZEROSKI; ZENKO, 2004)
Cascading	Pode ser considerado como um caso especial de *Stacking* principalmente devido à estrutura de aprendizagem em camadas. As predições são adicionadas ao conjunto de dados inicial e passadas para o próximo nível.	(GAMA; BRAZDIL, 2000)
Voting	São estratégias simples em que os resultados das decisões dos classificadores são calculados, por exemplo, tomando a classe que aparece na maioria dos casos. Pode-se adotar diretivas de ponderamento entre os votantes.	(LIN et al., 2003; KITTLER et al., 1998)

3.6 Aplicações em Biologia Computacional

Técnicas *ensemble* têm sido aplicadas como uma ferramenta eficaz para pesquisas de biomoléculas, principalmente no que se refere ao estudo de doenças (ROZZA et al., 2011). Nagi e Bhattacharyya (2013) realizaram um estudo empírico usando nove conjuntos de dados sobre câncer altamente dimensionais aplicados a três classificadores: J48 (QUINLAN, 1993; WITTEN; FRANK; HALL, 2011), NB (KOHAVI, 1996) e IBK (AHA; KIBLER; ALBERT, 1991). Em seguida, é apresentada uma metodologia denominada *SD-EnClass*, que visa combinar classificadores *ensemble* de diferentes famílias com base em uma estimativa simples do desempenho de classe de cada classificador individual. Os resultados mostraram que o modelo proposto melhora a precisão da classificação em comparação com a simples seleção do melhor classificador da combinação. Finalmente, o método proposto foi combinado com os resultados do uso de *Bagging*, *Boosting* e *Stacking*, no qual obteve resultados que foram significativamente melhores do que os classificadores *ensemble* individuais.

Na proposta de Eickholt e Cheng (2013) foi implementado um método baseado em *Boosting* para a predição de regiões desordenadas de proteínas a partir de sequências baseadas em conjuntos de redes profundas (do inglês, *deep networks*). Os resultados foram conseguidos, em parte, por um procedimento de reforço que é capaz de aumentar de forma constante a precisão e a área sob a curva ROC (do inglês, *Receiver Operating Characteristic*) ao longo de várias execuções do algoritmo.

Da mesma forma, pesquisas mais recentes também abordaram o uso de *Ensemble Learning* para identificar problemas tipicamente biológicos. A tese de Mendoza (2014) propõe uma solução ao problema de engenharia reversa de redes regulatórias genéticas a partir de dados pós-genômicos. A pesquisa busca investigar o uso de técnicas de *ensemble*

como forma de superar as limitações previstas e otimizar o processo de inferência através da exploração da diversidade de um conjunto de modelos. O *framework* desenvolvido implementa métodos computacionais, tanto para gerar redes diversificadas quanto para combinar essas predições em uma solução única.

No trabalho de Taghi et. al. (2014) foi utilizada uma técnica chamada *Select-Bagging*, que incorpora uma seleção de características a cada iteração do algoritmo *Bagging*, aplicado a uma série de conjuntos de dados de alta dimensionalidade. Também, Shi *et. al.* (2015) desenvolveram uma ferramenta para identificar sub-redes de interação de proteínas baseadas em uma estrutura de campo aleatório de *Markov* (BMRF) através da aplicação de *Bagging* a diferentes classificadores. Ao integrar dados de expressão gênica e dados de interação proteína-proteína (PPI) a ferramenta pode ser utilizada para identificar sub-redes biologicamente significativas.

Em Chen et. al. (2015), o *Stacking Multivariate Linear Regression* (SMLR) foi apresentado como um algoritmo que poderia desenvolver modelos preditivos para prever a capacidade de bioatividade de fitoterápicos a partir de suas impressões digitais cromatográficas. SMLR é um *meta-learner* que trabalha nos resultados dos *base-learners* constituintes (DZEROSKI; ZENKO, 2004). Também, Hu et. al. (2015) descrevem uma abordagem *Stacking* para identificar quadros de leitura abertos traduzidos dentro do mRNA, utilizando sequências de combinação relacionadas. O modelo de classificação foi desenvolvido com classes positivas e negativas sendo que o empilhamento foi utilizado para combinar os resultados de diferentes classificadores.

No trabalho de Zhang et al. (2016) é proposta uma estrutura *ensemble* que integra dados de expressão gênica e redes de interação proteína-proteína visando melhorar a acurácia de predição das medidas básicas de centralidade. A ideia dessa estrutura de conjunto é que diferentes interações proteína-proteína podem resultar em múltiplas contribuições para a essencialidade da proteína. Os experimentos consideraram cinco medidas padrão de centralidade: grau de centralidade, grau de intervalo, grau de proximidade, centralidade via matriz de adjacência e por sub-grafos. Os resultados mostraram que o *ensemble* proposto (uma combinação das técnicas de *Bagging* e *Boosting*) pôde melhorar consideravelmente a acurácia de predição das cinco medidas de centralidade individualmente.

4 FERRAMENTAS DE PREDIÇÃO ADOTADAS

As diferentes abordagens para induzir diversidade dentro de um sistema *ensemble* tem em comum a necessidade de se definir estratégias para combinar todas as soluções candidatas que compõem o conjunto em um único modelo de consenso (HANSEN; SALAMON, 1990). A metodologia proposta nesta dissertação é baseada na combinação dos resultados de diferentes ferramentas, assumindo-se o pressuposto de que a capacidade de generalização de um conjunto é frequentemente mais forte do que uma decisão individual (ZHOU, 2012). Para isso foram adotadas ferramentas descritas na literatura como capazes de predizer os efeitos na estabilidade de uma proteína sobre mutações pontuais através da variação da energia livre ($\Delta\Delta G$), ou seja, a diferença de energia livre entre uma proteína do tipo selvagem e o seu mutante.

É importante mencionar que o Capítulo 5 também aborda metodologias análogas a aqui proposta, porém, não fazem parte do *ensemble* aplicado pelo fato de não cumprirem algum dos pré-requisitos de implementação definidos. A escolha das ferramentas adotadas partiu das seguintes premissas:

- **Citações na literatura**: alguns dos trabalhos que antecederam a escrita desta dissertação abordaram a avaliação de ferramentas de predição do impacto de mutações encontradas na literatura bem como a discussão dos resultados obtidos. Isso permitiu a definição de uma lista de possíveis ferramentas a serem estudadas e o problema de pesquisa;

- **Predição numérica de** $\Delta\Delta G$: após a determinação da metodologia proposta, ficou estabelecido que o valor a ser resgatado das ferramentas de predição seria o $\Delta\Delta G$. Consequentemente, algumas delas não foram integradas pelo fato de predizer o resultado apenas de forma discreta (classes: Desestabilizante, Estabilizante; dentre outras classificações);

- **Tempo de processamento**: a predição do impacto de mutações pontuais abordada pela literatura destaca diferentes abordagens de implementação. Um importante fator de escolha nesse contexto, dado que a proposta desta dissertação visa o uso de múltiplas ferramentas, foi o tempo de predição do $\Delta\Delta G$. Com isso, as ferramentas

que apresentaram um tempo superior a 5 minutos em sua completa execução foram descartadas;

- **Possibilidade de automatização**: nenhuma das ferramentas estudadas apresentou uma API (*Application Programming Interface*). Devido a esse problema, a possibilidade de sistematização das tarefas também foi um fator decisório. É necessário mencionar que os detalhes de implementação de algumas ferramentas impossibilitaram a submissão automática de tarefas, por esse motivo, igualmente foram desconsideradas;

- **Licença de uso**: todas as ferramentas de predição utilizadas no *ensemble learning* desta dissertação são de uso gratuito, desde que sejam devidamente citados seus artigos originais.

4.1 *I-Mutant*

I-Mutant[1] é uma ferramenta baseada no algoritmo SVM (*Support Vector Machine*) (BOSER; GUYON; VAPNIK, 1992) na qual a predição de alterações de estabilidade de proteína em mutações pontuais são realizadas a partir da sua estrutura terciária ou sequência (CAPRIOTTI; FARISELLI; CASADIO, 2005). A ferramenta pode ser usada tanto como um classificador para predizer mudanças sobre a estabilidade de uma proteína mutante quanto para predizer os valores relacionados ao $\Delta\Delta G$ de dobramento (CAPRIOTTI; FARISELLI; CASADIO, 2005).

A metodologia adotada pelo *I-Mutant* utiliza como entrada um vetor de 42 posições. Os dois primeiros valores representam, respectivamente, a temperatura e o pH da proteína do tipo mutante. Os próximos valores são preenchidos pelos 20 tipos de aminoácidos existentes. Posteriormente, os últimos 20 valores de entrada codificam o ambiente do resíduo, sendo um "ambiente espacial" quando a estrutura da proteína está disponível ou os aminoácidos vizinhos, quando apenas a sequência da proteína é submetida (CAPRIOTTI; FARISELLI; CASADIO, 2005).

A Figura 15 exibe os parâmetros para a execução do *I-Mutant* na modalidade "*Protein Sructure*". Nela, é exigida a estrutura da proteína alvo (formato PDB) seguida da posição a ser substituída e o novo aminoácido. Também, é possível a parametrização da temperatura e do pH do experimento. Neste exemplo foi realizada a predição do impacto de uma mutação pontual na proteína de código PDB: 1STN, na qual realizou-se a substituição de uma treonina (T) por uma adenina (A) na posição 6. A saída é apresentada na Figura 16.

[1] http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi

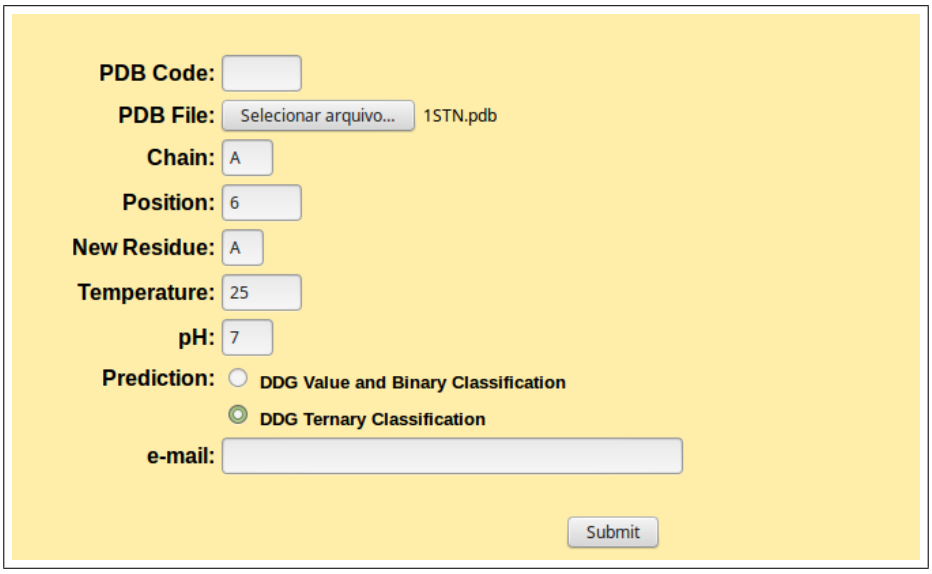


Figura 15: Exemplo de uso (submissão): *I-Mutant*.
Fonte: http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi

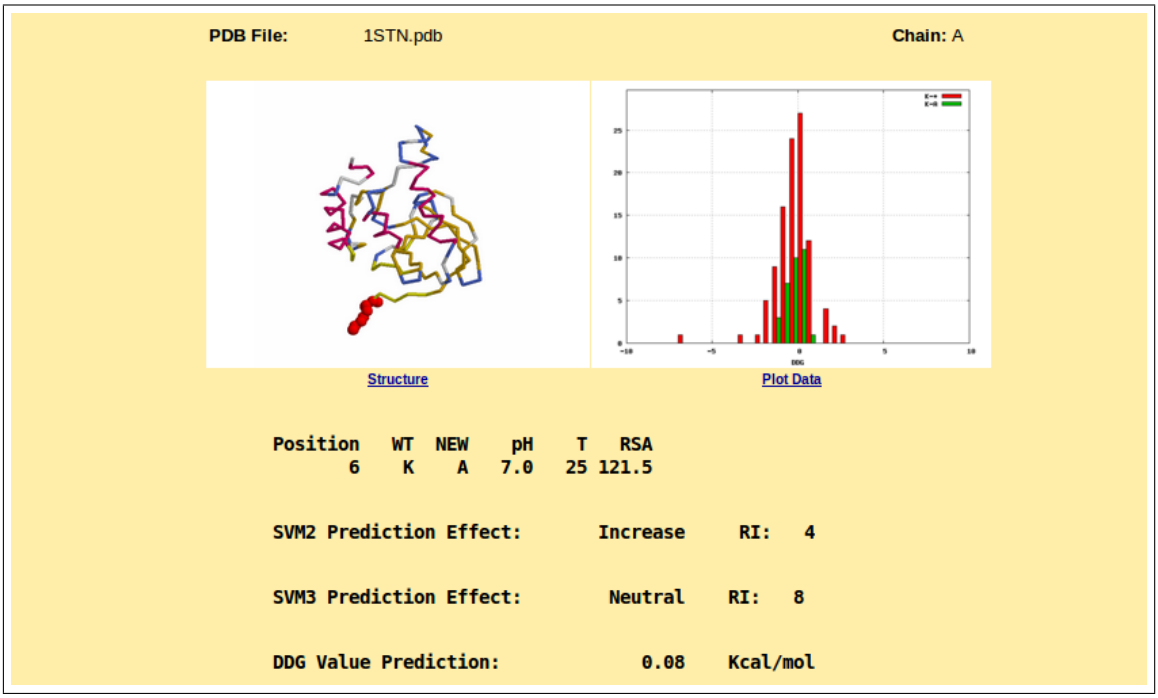


Figura 16: Exemplo de uso (saída): *I-Mutant*.
Fonte: http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi

Na modalidade de predição através da sequência da proteína, o *I-Mutant* utiliza múltiplos contextos de sequência de aminoácidos e os analisa em diferentes comprimentos, por exemplo: TIFQFPQDFMWGTATAAYQIE (21 aminoácidos), TIFQFPQDFMWG (12 aminoácidos), TIFQ (4 aminoácidos) (CAPRIOTTI; FARISELLI; CASADIO, 2005). Com isso, o SVM (*kernel* RBF) pode classificar as propriedades subjacentes do ambiente local de resíduos que conduzem à estabilidade/instabilidade da proteína em que ocorre a mutação. Essa descoberta indica que um pedaço de informação relevante para a estabilidade de dobramento da proteína pode ser rastreada até a sequência de vizinhos mais próximos do resíduo que sofre a mutação (CAPRIOTTI et al., 2008).

Do ponto de vista estrutural, a ferramenta considera como fator decisivo a centralidade dos resíduos das mutações em *C-alfa*, aplicados a crescentes valores de raio entre 0 Å e 12 Å (CAPRIOTTI et al., 2008). Dessa maneira, o valor da área acessível ao solvente pode ser calculado através do DSSP (KABSCH; SANDER, 1983), dividindo o valor da área de superfície acessível do resíduo mutado pela superfície de resíduo nativo, sendo que o resultado serve de entrada para o algoritmo SVM (CAPRIOTTI; FARISELLI; CASADIO, 2005).

Ao predizer os valores de $\Delta\Delta G$ associados as mutações, a correlação entre os valores experimentais e os esperados foi de: 0,71 (erro padrão de 1,30 Kcal/mol) com submissão de estrutura e 0,62 (erro padrão de 1,45 Kcal/mol) com submissão de sequência (CAPRIOTTI; FARISELLI; CASADIO, 2005). Os modelos do *I-Mutant* foram treinados e testados com um procedimento de validação cruzada em um conjunto de dados derivado da versão de dezembro de 2004 do *ProTherm*. Após um procedimento de filtragem, o conjunto de dados resultou em 2087 mutações pontuais distribuídas em 65 proteínas distintas. A última atualização do *I-Mutant* ocorreu no ano de 2006.

4.2 CUPSAT

CUPSAT[2] (*Cologne University Protein Stability Analysis Tool*) é uma ferramenta para a análise e predição de mudanças na estrutura de uma proteína causadas por mutações pontuais que se baseia principalmente nos ângulos de torção dos átomos que a compõem (PARTHIBAN; GROMIHA; SCHOMBURG, 2006). Dessa forma, a sua metodologia inclui a variação global da estabilidade da proteína alvo utilizando os potenciais dos átomos (cálculo da energia potencial de um sistema de átomos dadas suas posições no espaço) e a sua adaptação (favorável ou desfavorável) da combinação do ângulo de torção e o valor $\Delta\Delta G$ associado à predição de substituição *in-silico* (PARTHIBAN; GROMIHA; SCHOMBURG, 2006). Nesse ponto, baseados em Melo e Feytmans (1997), os átomos foram classificados em 40 tipos diferentes, de acordo com a sua localização, conectividade e natureza química. Da mesma forma, o modelo de predição pode distinguir o ambiente

[2] http://cupsat.tu-bs.de/

de aminoácidos através da sua acessibilidade ao solvente e a especificidade da estrutura secundária, utilizando a média das forças potencias para predizer a estabilidade da proteína (LAZAR et al., 2009). Semelhantemente, a ferramenta implementa a especificidade da estrutura secundária de mutações e potenciais de força média, seguida da classificação dos aminoácidos em hélices, folhas ou outros. Posteriormente, os aminoácidos pertencentes a cada um desses elementos de estrutura secundária são subdivididos de acordo com a sua acessibilidade ao solvente (PARTHIBAN; GROMIHA; SCHOMBURG, 2006).

A Figura 17 exibe a tela de submissão dos dados de entrada do CUPSAT. Neste exemplo foi utilizada a proteína de código PDB: 1LZ1, cadeia A, com a mutação pontual realizada pela troca de uma cisteína (C), posição 77, pelos demais aminoácidos. A saída do CUPSAT, exibida na Figura 18, consiste em informações sobre o impacto da mutação predita. Para isso são exibidas características estruturais (acessibilidade ao solvente, estrutura secundária e ângulos de torção) bem como informações sobre as alterações na estabilidade da proteína submetida para as 19 possíveis substituições de um único aminoácido.

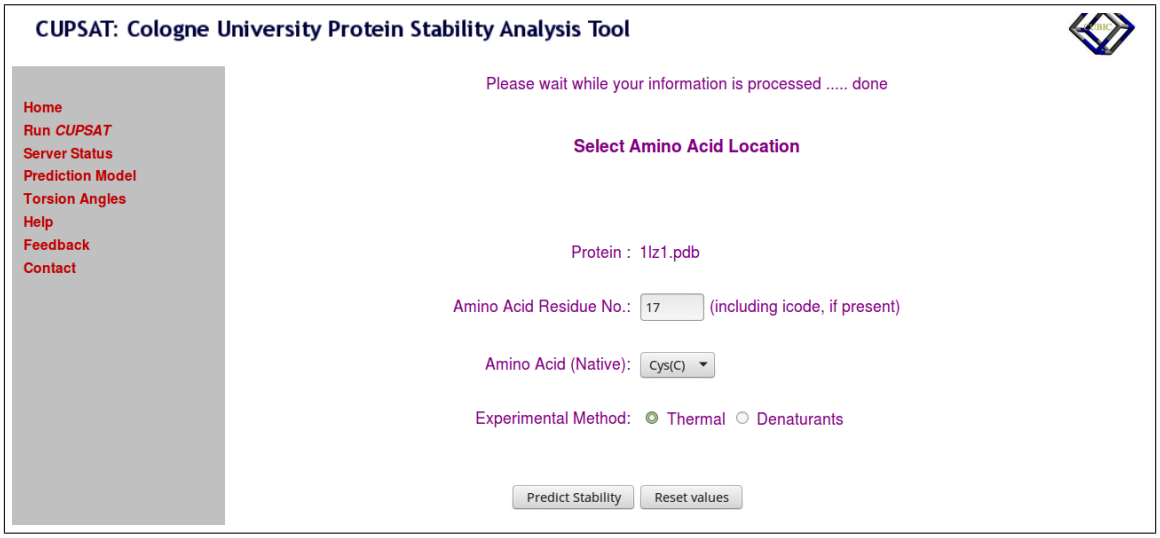
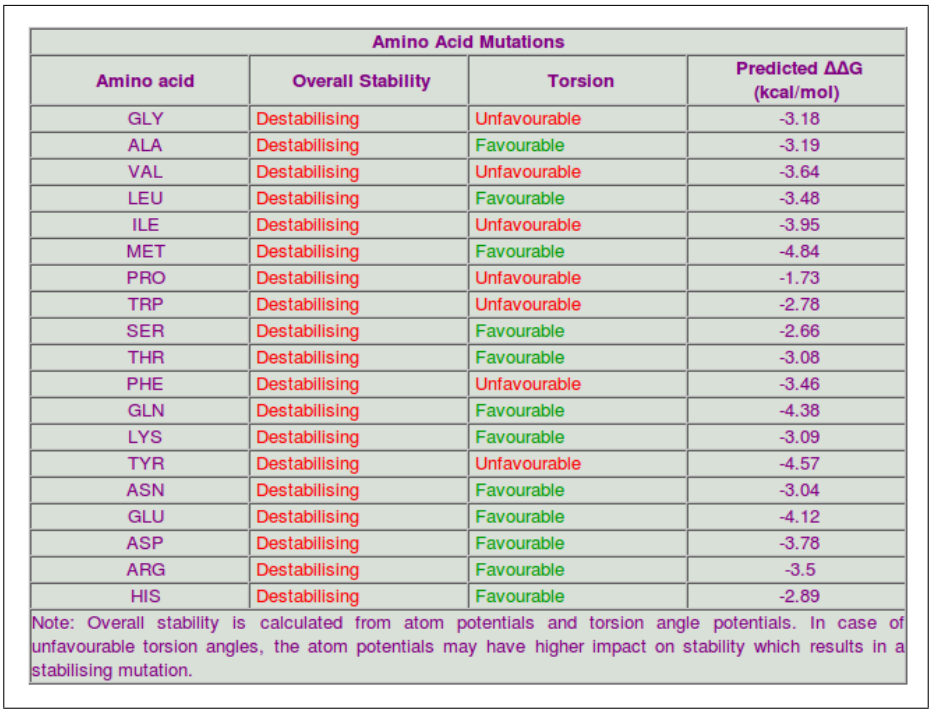


Figura 17: Exemplo de uso (submissão): CUPSAT.
Fonte: http://cupsat.tu-bs.de/cupsat/custompdb.htm

Os principais componentes que constroem o modelo preditivo do CUPSAT foram adquiridos com base em um conjunto de estruturas de proteínas não-redundantes obtido a partir da ferramenta PISCES (WANG; DUNBRACK, 2003). É importante ser mencionado que o CUPSAT não determina a sua predição baseada em algoritmos de aprendizado de máquina, assim como a maioria das ferramentas aqui abordadas. Em Saraboji, Gromiha e Ponnuswamy (2006) é descrito as diretivas aplicas na sua metodologia através dos conjuntos de dados utilizados, dentre elas podem ser destacadas: a frequência da ocorrência de mutantes e os efeitos físicos que contribuem para a diferença de estabilidade, para as tarefas de classificação; bem como a média de valores distribuídos em tabelas de mutantes, para os casos de predição numérica de $\Delta\Delta G$.



Amino Acid Mutations			
Amino acid	**Overall Stability**	**Torsion**	**Predicted ΔΔG (kcal/mol)**
GLY	Destabilising	Unfavourable	-3.18
ALA	Destabilising	Favourable	-3.19
VAL	Destabilising	Unfavourable	-3.64
LEU	Destabilising	Favourable	-3.48
ILE	Destabilising	Unfavourable	-3.95
MET	Destabilising	Favourable	-4.84
PRO	Destabilising	Unfavourable	-1.73
TRP	Destabilising	Unfavourable	-2.78
SER	Destabilising	Favourable	-2.66
THR	Destabilising	Favourable	-3.08
PHE	Destabilising	Unfavourable	-3.46
GLN	Destabilising	Favourable	-4.38
LYS	Destabilising	Favourable	-3.09
TYR	Destabilising	Unfavourable	-4.57
ASN	Destabilising	Favourable	-3.04
GLU	Destabilising	Favourable	-4.12
ASP	Destabilising	Favourable	-3.78
ARG	Destabilising	Favourable	-3.5
HIS	Destabilising	Favourable	-2.89

Note: Overall stability is calculated from atom potentials and torsion angle potentials. In case of unfavourable torsion angles, the atom potentials may have higher impact on stability which results in a stabilising mutation.

Figura 18: Exemplo de uso (saída): CUPSAT.
Fonte: http://cupsat.tu-bs.de/cupsat/custompdb.htm

Desse modo, para melhorar a predição e a sua especificidade, as mutações e os potenciais de força foram separados de acordo com regiões estruturais diferentes. Os critérios para medir a qualidade do modelo foram divididos em duas etapas principais: a capacidade de mostrar uma elevada correlação entre a predição realizada e o $\Delta\Delta G$ experimental das mutações selecionadas; assim como a capacidade de satisfazer diferentes testes de validação para provar a sua confiabilidade.

Os resultados do CUPSAT foram validados com 1538 mutações de desnaturação térmica e 1603 de desnaturação química, derivadas do *ProTherm* (versão do ano de 2006), de Topham, Srinivasan e Blundell (1997) assim como de Xu et al. (1998). Segundo seus autores, Parthiban, Gromiha e Schomburg (2006), os testes de amostra dividida (*split sample*) bem como de validação cruzada (*cross-validation*) mostraram um coeficiente de correlação máximo de 0,77, considerado os valores preditos e experimentais.

4.3 SDM

SDM[3] (*Site Directed Mutator*) é uma função de energia potencial utilizada por uma ferramenta *web* para calcular o escore de estabilidade relativo à diferença de energia livre ($\Delta\Delta G$) de proteínas do tipo selvagem e mutante (WORTH; PREISSNER; BLUNDELL, 2011). O algoritmo contido no SDM foi descrito pela primeira vez por Topham, Srinivasan e Blundell (1997) e usa frequências de substituição de aminoácidos específicos

[3]http://mordred.bioc.cam.ac.uk/ sdm/sdm.php

referentes às famílias de proteínas homólogas a submetida. Desta forma, duas pontuações de diferença de estabilidade são calculadas: uma usando substituições de aminoácidos considerando o seu ambiente e outra que avalia os aminoácidos predispostos (WORTH; PREISSNER; BLUNDELL, 2011).

Os parâmetros de predição do SDM consistem em um total de 14 atributos (9 de cadeias principais, 3 de acessibilidade ao solvente e 2 de ligações de hidrogênio) (WORTH; PREISSNER; BLUNDELL, 2011). A metodologia também inclui um conjunto baseado em tabelas de substituição de ambientes específicos com restrições conformacionais (do inglês, *Environment-Specific Substitution Tables*, ESSTs), cuja metodologia é descrita em Topham et al. (1993). A execução ilustrada na Figura 19 mostra os valores de entrada da ferramenta. Foi escolhida a proteína de código PDB: 3MBP, cadeia A, tendo a isoleucina (I) como aminoácido mutante na posição 345, que continha uma treonina. A saída é apresentada na Figura 20. Os resultados retornados incluem informações sobre o ambiente estrutural local do tipo selvagem e resíduos mutantes, um escore de predição de estabilidade, um arquivo no formato PDB e a predição do $\Delta\Delta G$.

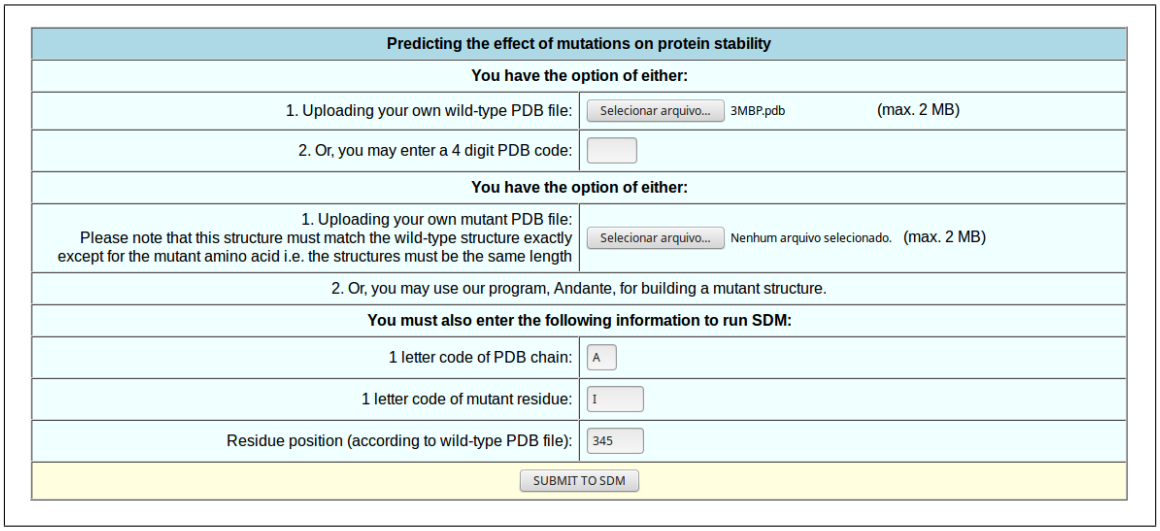


Figura 19: Exemplo de uso (submissão): SDM.
Fonte: http://mordred.bioc.cam.ac.uk/ sdm/sdm.php

Assim como no CUPSAT, a metodologia do SDM não inclui algoritmos de aprendizado de máquina tradicionais, pois é baseado em funções de energia potencial (TOPHAM; SRINIVASAN; BLUNDELL, 1997; WORTH; PREISSNER; BLUNDELL, 2011). O seu modelo preditivo obteve os dados a partir de alinhamentos de sequências de 371 proteínas da base de dados HOMSTRAD (MIZUGUCHI et al., 1998), consistindo, desse modo, em 1357 estruturas construídas utilizando uma versão modificada do programa *Makesub* (ainda não publicado) e um subconjunto de mutações derivadas do *ProTherm* (versões do anos de 2004 e 2009) (WORTH; PREISSNER; BLUNDELL, 2011).

Figura 20: Exemplo de uso (saída): SDM.
Fonte: http://mordred.bioc.cam.ac.uk/ sdm/sdm.php

O SDM foi previamente validado por seus autores, Worth, Preissner, Blundell (2011), através de um conjunto de dados com 230 mutações pontuais publicadas por Dehouck et al. (2009), atingindo 74% de acurácia na predição de estabilidade em mutações pontuais bem como um coeficiente de correlação de 0,60 do valor de $\Delta\Delta G$, ambos comparados com os resultados experimentais. Na versão atual da ferramenta, atualizada em 2011, foram acrescentados dados de substituição através de famílias de proteínas adicionais.

4.4 mCSM

mCSM[4] (*mutation Cutoff Scanning Matrix*) é uma ferramenta que utiliza o conceito de assinaturas estruturais a fim de estudar e predizer o impacto de mutações pontuais em proteínas (PIRES; ASCHER; BLUNDELL, 2014a). O mCSM é uma versão continuada de dois métodos: CSM (*Cutoff Scanning Matrix*), uma abordagem para tarefas de classificação estrutural e predição de função de proteína baseada em grafos (PIRES et al., 2011); e aCSM (*atomic level Cutoff Scanning Matrix*), uma extensão a nível atômico das assinaturas estruturais aplicadas aos *binding pockets* (em português, "bolsos de ligação"), um provável lugar para um sítio de ligação na estrutura (PIRES et al., 2013). O *workflow* da ferramenta é dividido em três componentes principais, segundo Pires, Ascher e Blundell (2014a):

- **Grafos baseados em padrões de distância dos átomos**: são usados 3 tipos de classificação de átomos, sendo uma classificação simples na qual não há distinção entre os átomos, uma classificação binária em que os átomos são marcados como polar ou hidrofóbico e uma classificação farmacofórica que divide os átomos em categorias, dentre elas, hidrofóbico, positivo, negativo, aromático e neutro;

[4] http://bleoberis.bioc.cam.ac.uk/mcsm/stability

- **Mudanças farmacóforas**: a frequência de cada tipo de farmacóforo em um resíduo é apresentada como um vetor P, sendo que a diferença $P_{mudança}$ entre a contagem de farmacóforo para o resíduo mutante (P_{mut}) e do tipo selvagem (P_{sel}) é calculada ($P_{mudança}$ = P_{mut} - P_{sel}) e anexada à assinatura;

- **Condições experimentais**: condições nas quais os dados termodinâmicos são coletados, tais como pH, temperatura e acessibilidade ao solvente.

A opção escolhida para a execução do mCSM foi o "*Single mutation*" (Figura 21) sendo que a saída é exibida na Figura 22. Nesse exemplo, foi utilizada a proteína de código PDB: 4LYZ, resíduo de posição 13 e substituição de uma lisina (K) por um aspartato (D). Por esse motivo, no campo *Mutation* é informado o valor K13D.

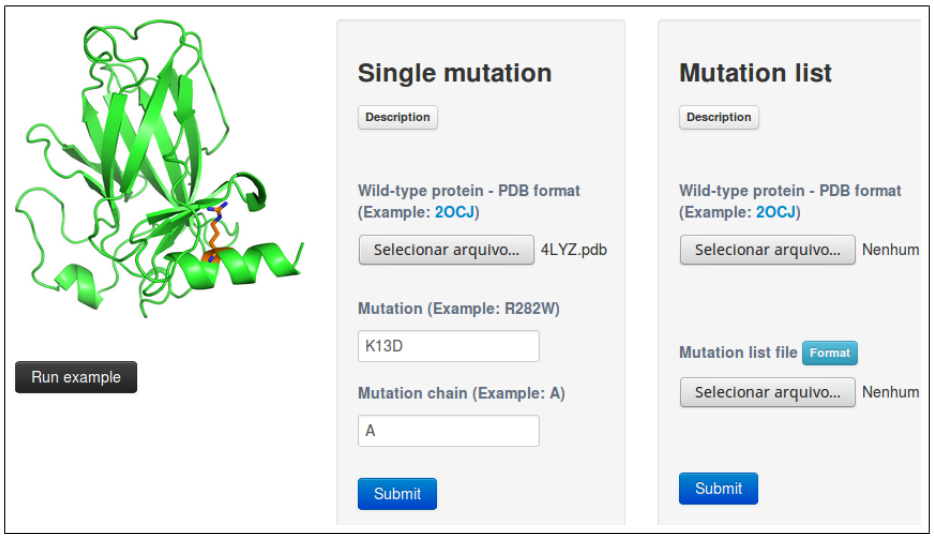


Figura 21: Exemplo de uso (submissão): mCSM.
Fonte: http://bleoberis.bioc.cam.ac.uk/mcsm/stability

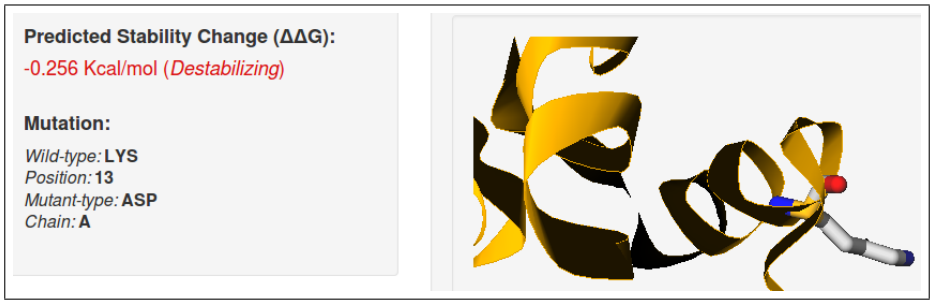


Figura 22: Exemplo de uso (saída): mCSM.
Fonte: http://bleoberis.bioc.cam.ac.uk/mcsm/stability

Alguns dos dados experimentais que amparam a construção do modelo preditivo do mCSM foram obtidos do *ProTherm* e outros do SKEMPI (*Structual database of Kinetics and Energetics of Mutant Protein Interactions*) (MOAL; FERNÁNDEZ-RECIO, 2012),

uma base de dados que descreve complexos proteína-proteína. Com o intuito de compreender o papel das mutações em doenças, os autores do mCSM, Pires and Ascher e Blundell (2014a), abordam o impacto de mutações pontuais sobre a estabilidade da proteína, assim como interações proteína-proteína e proteína-ácido-nucleico (PIRES; ASCHER; BLUNDELL, 2014a). As assinaturas mCSM foram utilizadas para treinar um modelo de regressão que obteve uma correlação (valores preditos e experimentais) de 0,82 e um erro padrão de 1,02 Kcal/mol. Os resultados foram validados em um teste que envolveu o uso da ferramenta para predizer as mudanças de estabilidade em 42 mutações que ocorrem na proteína supressora tumoral *p53* (código PDB: 2OCJ). Segundo Beroud e Soussi (2003), mais de 50% dos cânceres humanos são portadores de mutações com perda de função no fator de transcrição *p53*. Nesse estudo de caso foi identificado a mutação de uma arginina para um triptofano na posição 282 tendo efeitos significativos sobre a estabilidade da proteína de código PDB: 2OCJ (PIRES; ASCHER; BLUNDELL, 2014a).

4.5 DUET

DUET[5] é um servidor *web* que utiliza uma metodologia *ensemble* a fim de estudar o efeito de mutações pontuais em proteínas. A ferramenta consolida duas abordagens complementares (SDM e mCSM) em uma predição baseada em um consenso proveniente da combinação dos resultados obtidos pelos dois métodos (PIRES; ASCHER; BLUNDELL, 2014b). Dessa maneira, informações complementares sobre a mutação, tal como a estrutura secundária (utilizada pelo SDM) e um vetor de farmacóforico que representa as alterações entre o tipo selvagem e o resíduo mutante (utilizado pelo mCSM), também são calculadas e usadas pelo DUET (PIRES; ASCHER; BLUNDELL, 2014b). Com isso, dada uma mutação pontual em uma estrutura de proteína, a ferramenta realiza o *ensemble learning* combinando os dois métodos integrados de uma forma não linear utilizando SVR (*Support Vector Regression*) com um *kernel* RBF (SCHOLKOPF et al., 1997).

A execução ilustrada na Figura 23 apresenta os parâmetros de entrada do DUET, sendo escolhida para predição a proteína de código PDB: 2HPR, cadeia A, substituindo uma glicina por uma lisina na posição 49. A saída é apresentada na Figura 24. Como o DUET utiliza uma abordagem integrada, também é exibido os resultados das demais ferramentas SDM e mCSM. Tais resultados preditos são expressos através da variação na energia livre de Gibbs ($\Delta\Delta G$), sendo que valores negativos indicam mutações desestabilizantes.

Para a validação do DUET, o modelo foi treinado através de um conjunto de dados de mutações com valores termodinâmicos experimentais derivados da base de dados *ProTherm*. Consequentemente, foi obtido um coeficiente de correlação de 0,74 (valores preditos e experimentais) durante o treinamento e 0,71 para um conjunto de teste, bem como 0,82 e 0,79, respectivamente, após a remoção de valores atípicos (*outliers*) (PIRES;

[5]http://bleoberis.bioc.cam.ac.uk/duet/stability

ASCHER; BLUNDELL, 2014b). Para minimizar o risco de *overfitting*, dois testes cegos foram executados para validar os experimentos. O primeiro conjunto de dados foi composto de 351 mutações não redundantes no nível de posição de aminoácido a ser substituído. Diferentemente, o segundo conjunto de dados conteve 42 mutações dentro do domínio de ligação do DNA da proteína supressora de tumor *p53*, com seus efeitos termodinâmicos caracterizados experimentalmente e disponíveis na literatura. Nenhuma dessas mutações validadas estava presente no conjunto de treinamento (PIRES; ASCHER; BLUNDELL, 2014b).

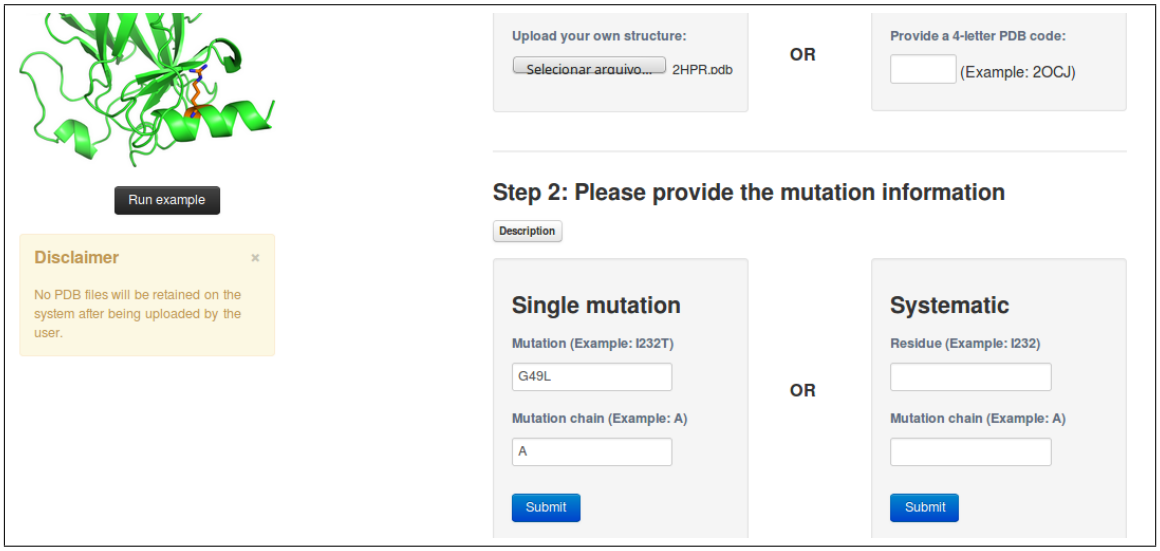


Figura 23: Exemplo de uso (submissão): DUET.
Fonte: http://bleoberis.bioc.cam.ac.uk/duet/stability

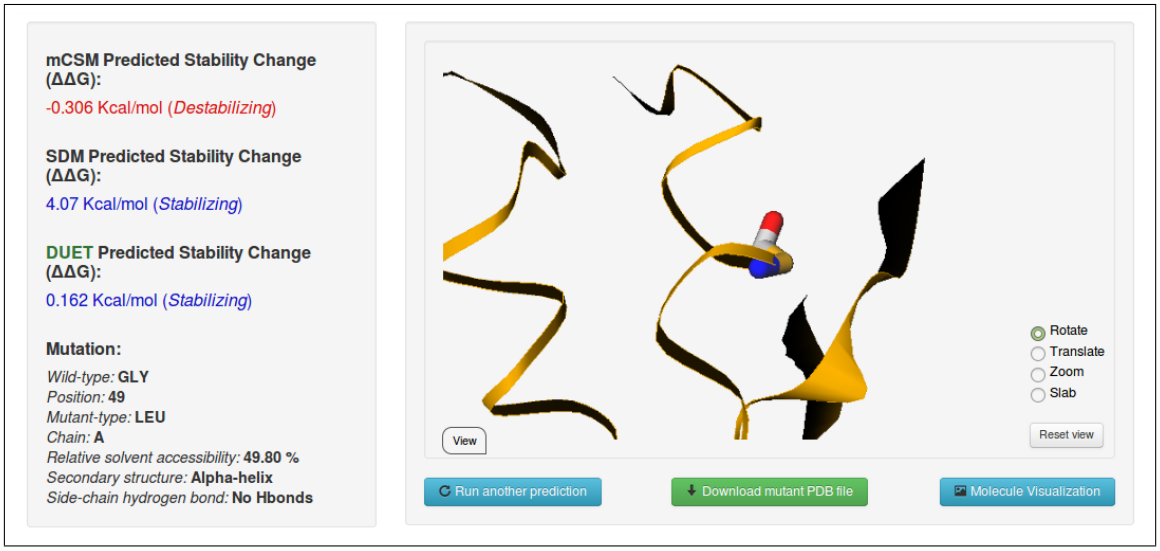


Figura 24: Exemplo de uso (saída): DUET.
Fonte: http://bleoberis.bioc.cam.ac.uk/duet/stability

4.6 iRDP

O iRDP[6] (*in-silico Rational Designing of Proteins*) é uma plataforma *web* unificada que compreende três módulos: iCAPS (*in-silico Comparative Analysis of Protein Structures*), *iStability* e *iMutants*. Cada módulo aborda diferentes particularidades de engenharia de proteínas destinada à estabilidade melhorada (PANIGRAHI et al., 2015). Enquanto o iCAPS ajuda na seleção da proteína alvo com base em fatores que contribuem para a estabilidade estrutural, o *iStability* oferece abordagens *in-silico* baseadas em proteínas conhecidas, contribuindo, dessa maneira, para a identificação e predição da estabilidade em potenciais locais de mutação. O *iMutants* visa avaliar proteínas mutantes com base em mudanças na rede de interação local e no grau de conservação dos resíduos das regiões de mutação (PANIGRAHI et al., 2015). No entanto, o *iStability*, principal módulo, usa a abordagem DSSP (*Dictionary of protein Secondary Structure*) (KABSCH; SANDER, 1983) para detectar as estruturas secundárias das proteínas de entrada e o NACCESS (HUBBARD; THORNTON, 1993) para uma estimativa de áreas de superfície acessível ao solvente. Complementando a tarefa de predição, a ferramenta *FindGeo* (ANDREINI; CAVALLARO; LORENZINI, 2012) é utilizada para a análise de locais de ligação e geometria, das quais a energia livre *Gibbs* de dobramento é calculada usando o *FoldX* (SCHYMKOWITZ et al., 2005).

Atualmente, o iRDP utiliza métodos baseados tanto em funções de energia potencial empíricas quanto em algoritmos de aprendizado de máquina, como SVMs e Redes Neurais. Uma vez que esses métodos são fundamentados em diferentes conjuntos de dados, a predição de mutações que são distantes para o conjunto de dados de treinamento se torna mais eficaz (PANIGRAHI et al., 2015). A utilização da ferramenta ilustrada na Figura 25 apresenta os parâmetros de entrada do iRDP, sendo escolhida a proteína de código PDB: 1KA6, cadeia A, e a substituição de uma fenilalanina por uma serina como aminoácido mutante na posição 87. A saída é apresentada na Figura 26.

A abordagem de análise de interação local do *iStability* foi estendida para analisar mutações experimentalmente validadas e listadas na base de dados *ProTherm*. As mutações também foram fornecidas na forma de iATMs (*in-silico Analysis of Thermally stable Mutants*), uma metodologia que as organiza em três seções com base no tipo de mutação (única, dupla ou múltipla), sendo um recurso de informação suplementar para o *ProTherm*. Na validação do *iStability*, um total de 81 mutações foram selecionadas a partir de um conjunto de 40 estruturas de proteínas, na qual foi obtida uma acurácia de predição de 84% (em relação aos valores experimentais).

[6] http://irdp.ncl.res.in/index.html

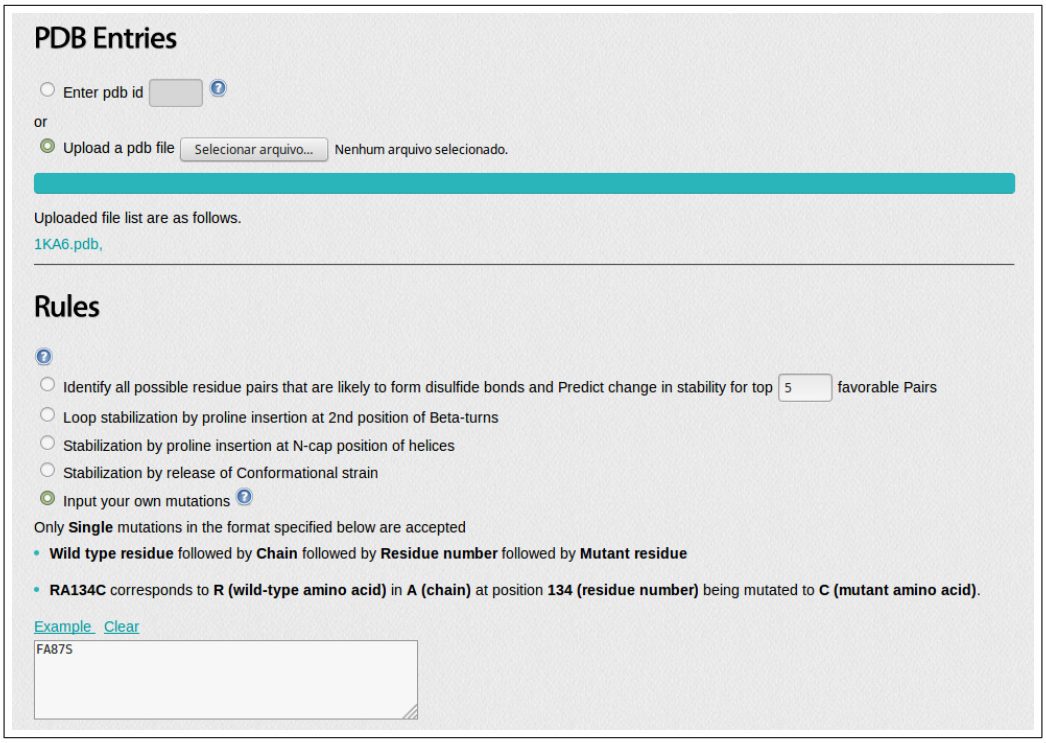
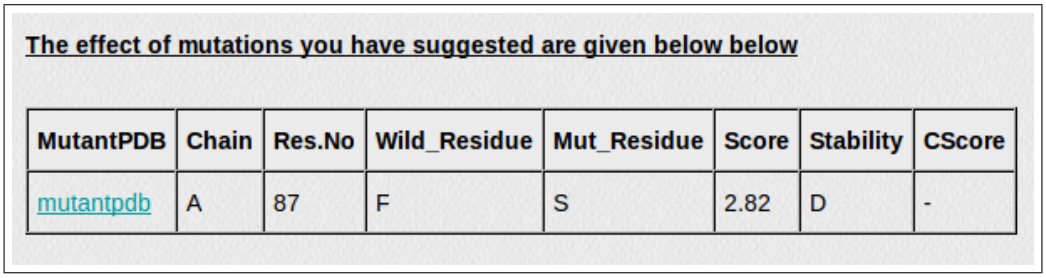


Figura 25: Exemplo de uso (submissão): iRDP.
Fonte: http://irdp.ncl.res.in/Proteng.html



The effect of mutations you have suggested are given below below

MutantPDB	Chain	Res.No	Wild_Residue	Mut_Residue	Score	Stability	CScore
mutantpdb	A	87	F	S	2.82	D	-

Figura 26: Exemplo de uso (saída): iRDP.
Fonte: http://irdp.ncl.res.in/Proteng.html

4.7 MAESTRO

MAESTRO[7] (*Multi AgEnt STability pRedictiOn*) é uma ferramenta de predição do impacto de mutações pontuais baseada em funções estatísticas de pontuação (do inglês, *Statistical Scoring Functions*) (SSFs), combinada a abordagens de AM (LAIMER et al., 2016). As SSFs são complementadas por uma seleção de propriedades globais e locais da proteína. Como propriedade global é utilizado o tamanho da molécula. O ambiente local da mutação é descrito pelo estado da estrutura secundária bem como a área de superfície acessível ao solvente (LAIMER et al., 2015). A predição de estrutura secundária abordada pelo MAESTRO é realizada com uma versão aperfeiçoada do algoritmo *Saba* (PARK et al., 2011) e uma adaptação da biblioteca *Geometry* (VOSS; GERSTEIN, 2005).

O MAESTRO realiza uma predição *ensemble* dividida em três partes principais: um escore baseado em SSFs e nos demais valores de entrada; uma predição realizada por agentes; e o cálculo de consenso entre esses mesmos agentes (LAIMER et al., 2015). Para isso, a metodologia emprega o SVM, através da *LIBSVM* (CHANG; LIN, 2011), complementada a um SVR, configurado com um *kernel* Gaussiano, sendo que os parâmetros de predição dos múltiplos agentes de regressão linear são calculados através de uma implementação fornecida pela *GNU Scientific Library* (GOUGH, 2009).

A Figura 27 apresenta os parâmetros de entrada para a execução do MAESTRO em uma predição de mutação pontual na proteína de código PDB: 2DRI, cadeia A, realizando-se a substituição de uma leucina por uma cisteína na posição 62. A saída é exibida na Figura 28.

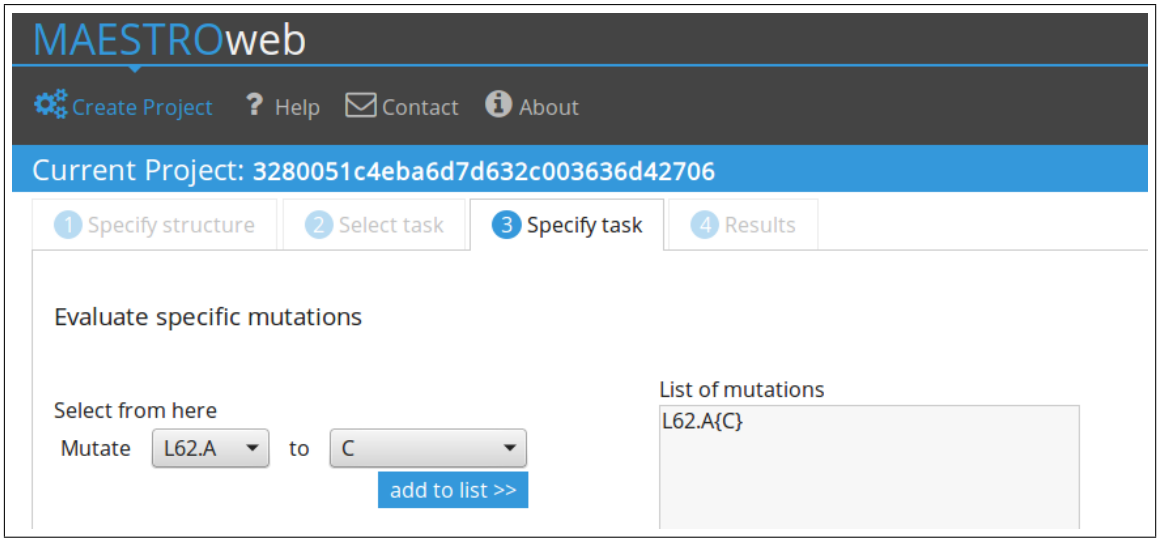


Figura 27: Exemplo de uso (submissão): MAESTRO.
Fonte: https://che.sbg.ac.at/maestro/web/maestro/workflow

[7]https://biwww.che.sbg.ac.at/maestro/web

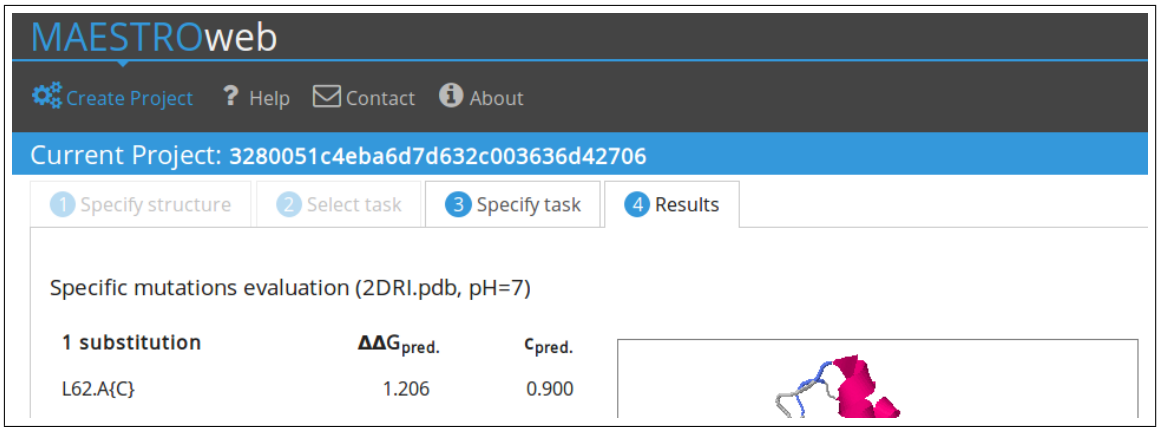


Figura 28: Exemplo de uso (saída): MAESTRO.
Fonte: https://che.sbg.ac.at/maestro/web/maestro/workflow

A ideia básica da abordagem de predição multiagente é melhorar o poder de acurácia em relação aos métodos de aprendizado convencionais, limitando, assim, o risco de *outliers* e *overfitting*. Os autores do MAESTRO avaliaram a ferramenta em cinco diferentes conjuntos de dados derivados do *ProTherm*, alcançando coeficientes de correlação (valores preditos e experimentais) entre 0,7 e 0,8 para as instâncias submetidas.

4.8 Comparativo entre as ferramentas de predição

As ferramentas apresentadas neste capítulo tiveram seus modelos preditivos treinados em diferentes condições e com conjuntos de dados distintos. Tais ferramentas abordaram a predição dos valores reais de $\Delta\Delta G$ e uma classificação baseada em duas classes: desestabilizante ou estabilizante; bem como em preditores de três classes, na qual a mutação é classificada como: desestabilizante, neutra ou estabilizante. Por razões práticas, a maioria dos métodos de predição faz uso de representações simplificadas da proteína, limitando, assim, o número de conformações a serem avaliadas (o chamado espaço conformacional) com a adoção de funções de energia empíricas (ou semi-empíricas) ou baseadas em conhecimento (*knowledge-based*), as quais capturam as forças mais importantes que impulsionam e estabilizam o enovelamento da proteína (VERLI et al., 2014).

Algumas ferramentas, como por exemplo, o *I-Mutant* e o mCSM, foram criadas com a ajuda de técnicas de AM em que exemplos de proteínas (tipo selvagem e mutante) têm suas medidas termodinâmicas de $\Delta\Delta G$ determinadas experimentalmente e disponíveis em bancos de dados biológicos. Outras ferramentas, como o MAESTRO e o DUET, podem combinar técnicas de análise estatística e metodologias *ensemble*. É importante mencionar que todas as ferramentas abordadas neste capítulo tem a sua execução via *web* e estão em funcionamento até o corrente ano de 2017. As Tabelas 4, 5 e 6 resumem as características apresentadas por cada uma das ferramentas de predição elencadas.

Tabela 4: Comparativo entre as ferramentas de predição: dados de publicação.

Ferramenta	Ano	Referência
I-Mutant	2005	(CAPRIOTTI; FARISELLI; CASADIO, 2005)
CUPSAT	2006	(PARTHIBAN; GROMIHA; SCHOMBURG, 2006)
SDM	2011	(WORTH; PREISSNER; BLUNDELL, 2011)
mCSM	2014	(PIRES; ASCHER; BLUNDELL, 2014a)
DUET	2014	(PIRES; ASCHER; BLUNDELL, 2014b)
iRDP	2015	(PANIGRAHI et al., 2015)
MAESTRO	2016	(LAIMER et al., 2016)

Tabela 5: Comparativo entre as ferramentas de predição: dados de implementação.

Ferramenta	Conjunto de dados	Tipo de entrada	Tipo de saída
I-Mutant	1948 mutações pontuais	sequência ou estrutura	$\Delta\Delta G$ (3 classes)
CUPSAT	3141 mutações pontuais	estrutura	$\Delta\Delta G$ (2 classes)
SDM	2686 mutações pontuais	estrutura	$\Delta\Delta G$ (2 classes)
mCSM	4965 mutações pontuais	estrutura	$\Delta\Delta G$ (3 classes)
DUET	2848 mutações pontuais	estrutura	$\Delta\Delta G$ (3 classes)
iRDP	4427 mutações pontuais	estrutura	$\Delta\Delta G$ (2 classes)
MAESTRO	4573 mutações pontuais	estrutura	$\Delta\Delta G$ (2 classes)

Tabela 6: Comparativo entre as ferramentas de predição: principais características.

Ferramenta	Principais características	Utiliza *ensemble*?
I-Mutant	Predição baseada em SVM, sendo realizada a partir da sequência ou estrutura da proteína.	Não
CUPSAT	Utiliza potenciais dos átomos e seus ângulos de torção para predizer o $\Delta\Delta G$ de uma mutação.	Não
SDM	Implementa uma função de energia potencial que usa as frequências de substituição de aminoácidos para calcular o escore de estabilidade.	Não
mCSM	Utiliza assinaturas de predição baseadas em grafos que codificam as distâncias entre os átomos.	Não
DUET	Integra duas abordagens complementares em uma predição via consenso obtida por combinação de resultados e uma tarefa de classificação.	Sim
iRDP	Fornece uma plataforma unificada que compreende três módulos destinados à engenharia de proteínas e análise de mutações pontuais.	Não
MAESTRO	Implementa um sistema de AM multiagente com decisão via consenso e uma estimativa de confiança da predição correspondente.	Sim

5 TRABALHOS RELACIONADOS

Este capítulo apresenta alguns trabalhos correlatos ao tema proposto nesta dissertação através do desenvolvimento de ferramentas de predição do impacto de mutações pontuais em proteínas.

5.1 Chen, Lin e Chu (2013)

No trabalho intitulado, "*iStable: off-the-shelf predictor integration for predicting protein stability changes*", de Chen, Lin e Chu (2013), é proposta uma ferramenta de predição do impacto de mutações pontuais que integra um classificador SVM a uma abordagem baseada em resultados de outras ferramentas correlatas. Os resultados apresentados pelos autores são exibidos separadamente nas Tabelas 7 e 8.

Tabela 7: Resultados de predição do conjunto de dados M1311 (1311 mutações).

Ferramenta	Acurácia
I-Mutant	80,00%
AUTO-MUTE	95,08%
MUpro	89,06%
CUPSAT	74,02%
iStable	**96,90%**

Fonte: Adaptado de (CHEN; LIN; CHU, 2013).

Tabela 8: Resultados de predição do conjunto de dados M1820 (1820 mutações).

Ferramenta	Acurácia
I-Mutant	67,00%
AUTO-MUTE	70,00%
MUpro	68,20%
PoPMuSiC	73,60%
CUPSAT	62,80%
iStable	**75,20%**

Fonte: Adaptado de (CHEN; LIN; CHU, 2013).

Na construção do *iStable*, cinco ferramentas de predição de foram escolhidas como votantes, sendo elas: *I-Mutant* (CAPRIOTTI; FARISELLI; CASADIO, 2005), *MUpro* (CHENG; RANDALL; BALDI, 2006), CUPSAT (PARTHIBAN; GROMIHA; SCHOMBURG, 2006), *PoPMuSiC* (DEHOUCK et al., 2009) e AUTO-MUTE (MASSO; VAISMAN, 2010). Em ambos os conjuntos de dados, o *iStable* mostrou uma melhor acurácia quando comparado aos demais preditores. A técnica *ensemble* de voto majoritário é adotada como elemento principal na definição do resultado em conjunto. A etapa seguinte da ferramenta consiste no uso de informações de sequência de proteínas classificadas por famílias, formando, desse modo, o arquivo de entrada para o treinamento via SVM. Após a submissão de uma tarefa pelo usuário, o *iStable* determina se a mutação é estabilizante ou desestabilizante, reunindo a decisão do modelo bem como o consenso das ferramentas participantes do *ensemble*. Para a definição do modelo preditivo do *iStable* foram obtidos resultados de predição das ferramentas integradas.

5.2 Malinka (2015)

No trabalho proposto por Malinka (2015) é apresentada uma meta-ferramenta para a detecção de alterações na estabilidade de proteínas através de mutações de aminoácidos, tendo como objetivo principal a criação de uma metodologia que combine sete ferramentas de predição já estabelecidas: *I-Mutant* (CAPRIOTTI; FARISELLI; CASADIO, 2005), CUPSAT (PARTHIBAN; GROMIHA; SCHOMBURG, 2006), iPTREE-STAB (HUANG; GROMIHA; HO, 2007), SDM (WORTH et al., 2007), *PoPMuSiC* (DEHOUCK et al., 2009), AUTO-MUTE (MASSO; VAISMAN, 2010) e mCSM (PIRES; ASCHER; BLUNDELL, 2014a). Para isso, o modelo preditivo foi induzido por diferentes métodos de AM que suportam regressão, no qual foram avaliados 28 algoritmos de AM que permitem prever variáveis contínuas, ou seja, os valores de $\Delta\Delta G$. A Tabela 9 exibe o coeficiente de correlação (valores preditos e experimentais) para o conjunto de dados definido pelos autores e aplicado as ferramentas correlatas bem como ao algoritmo *KStar*, que obteve os melhores resultados com a metodologia proposta.

Tabela 9: Resumo preditivo dos resultados obtidos por Malinka (2015).

Ferramenta	Coeficiente de correlação
AUTO-MUTE	0,58
SDM	0,36
CUPSAT	0,17
I-Mutant	0,52
iPTREE-STAB	0,5
mCSM	0,48
KStar	**0,71**

Fonte: Adaptado de (MALINKA, 2015).

Dentre os 8 melhores algoritmos testados (*Majority*, *Gaussian Process*, *LIBSVM* , *KStar*, *M5Rules*, *M5P*, *Bagging*, *Random SubSpace*), o *KStar* obteve a melhor correlação para o conjunto de dados de adotado, composto por 416 mutações estabilizantes e 1179 desestabilizantes, ambas obtidas do *ProTherm*. No entanto, uma das principais desvantagens do método proposto por Malinka (2015) é o fato de estar descrito somente na forma de um *workflow*, não oferecendo uma interface de uso ao especialista, em potencial, os biólogos.

5.3 Fariselli et al. (2015)

Em Fariselli et al. (2015) foi descrito o INPS (*a predictor of the Impact of Non-synonymous-variations on Protein Stability*), uma ferramenta que computa os valores de $\Delta\Delta G$ de variantes de proteínas sem requerer o conhecimento da estrutura proteica (utiliza a sequência). A metodologia abordada pela ferramenta para treinar os modelos preditivos inclui a reversibilidade termodinâmica das mutações, ou seja, é considerado a variação inversa em uma proteína (por exemplo, troca de aminoácidos GA e AG), sendo caracterizada pelo valor negativo de $\Delta\Delta G$ determinado experimentalmente. Com isso, é possível reformular as propriedades termodinâmicas do problema ($\Delta\Delta G\ (A, B)$ = -$\Delta\Delta G\ (B, A)$), bem como equilibrar a distribuição das medidas experimentais de mudanças de energia livre disponíveis.

A Tabela 10 exibe a correlação entre os valores experimentais e preditos quando aplicados a um conjunto de teste específico da proteína *p53* (42 mutações pontuais). É importante destacar que na abordagem *ensemble* na qual a predição do INPS (sequencial) é combinada com a do mCSM (estrutural) foi atingido o melhor desempenho no conjunto de dados *p53*.

Tabela 10: Resultados de predição do conjunto de dados *p53* (42 mutações).

Ferramenta	Coeficiente de correlação
mCSM	0,68
PoPMuSiC	0,56
DUET	0,68
INPS	**0,71**
INPS + mCSM	**0,75**

Fonte: Adaptado de (FARISELLI et al., 2015).

Posteriormente, a Tabela 11 apresenta os resultados preditos através do conjunto de dados S2468 (2468 mutações pontuais), nesse caso, utilizando a técnica de validação cruzada (*5 folds*).

Tabela 11: Resultados de predição do conjunto de dados S2468 (2468 mutações).

Ferramenta	Coeficiente de correlação
mCSM	0,69
PoPMuSiC	0,63
INPS	**0,60**

Fonte: Adaptado de (FARISELLI et al., 2015).

O preditor INPS consiste em um SVR treinado no conjunto de dados S2648 (2468 mutações pontuais) (PIRES; ASCHER; BLUNDELL, 2014a) utilizando sete características divididas em duas abordagens: seis descritores que codificam o tipo de mutação e um descritor que codifica a informação evolutiva. A informação evolutiva é derivada analisando múltiplas sequências alinhadas com sequências obtidas executando a ferramenta JackHMMer (EDDY, 2011) aplicada ao banco de dados UNIREF90, versão de setembro de 2014 (APWEILER et al., 2004).

5.4 Witvliet et al. (2016)

No trabalho publicado por Witvliet et al. (2016), denominado "*ELASPIC web-server: proteome-wide structure based prediction of mutation effects on protein stability and binding affinity*", é proposta uma abordagem de *ensemble* para a predição dos efeitos de mutações em dobramento de proteínas e interações proteína-proteína. O ELASPIC (*Ensemble Learning Approach for Stability Prediction of Interface and Core mutations*) pode ser usado para avaliar o efeito de mutações em qualquer proteína do banco de dados *UniProt* (APWEILER et al., 2004) além do *Protein Data Bank* (BERMAN et al., 2000).

Os resultados exibidos nas Tabelas 12 e 13 mostram um coeficiente de correlação atingindo pelo ELASPIC entre os valores de $\Delta\Delta G$ experimentais e preditos de 0,77 e 0,75, para os conjuntos de dados do *ProTherm* e SKEMPI, respectivamente. Os resultados foram igualmente comparados com de outras ferramentas correlatas.

Tabela 12: Resultados de predição do conjunto de dados *ProTherm* (3463 mutações).

Ferramenta	Coeficiente de correlação
ProMaya	0,74
Prethermut	0,71
PoPMuSiC	0,62
FoldX	0,51
ELASPIC	**0,77**

Fonte: Adaptado de (BERLINER et al., 2014).

Tabela 13: Resultados de predição do conjunto de dados SKEMPI (857 mutações).

Ferramenta	Coeficiente de correlação
BeAtMuSiC	0,68
FoldX	0,44
ELASPIC	**0,75**

Fonte: Adaptado de (BERLINER et al., 2014).

Para construir os modelos de homologia de domínios e de interações domínio-domínio, o ELASPIC utiliza o *Modeller* (FISER; SALI, 2003) com otimizações feitas através do *FoldX* (SCHYMKOWITZ et al., 2005). Dessa forma, é possível induzir mutações através de escores de energia baseados em sequências e outras características relevantes a fim de predizer o impacto termodinâmico de uma mutação na estabilidade de um único domínio ou na afinidade entre dois domínios (proteína-proteína).

O ELASPIC foi treinado e testado com 3463 mutações depositadas nos bancos de dados *ProTherm* e SKEMPI. Para a tarefa de AM, a ferramenta adotou o uso do SVM em conjunto com o algoritmo SGB-DT (*Stochastic Gradient Boosting of Decision Trees*), utilizado para ajustar uma função não-linear com o intuito de minimizar o erros de predição. O modelo preditivo foi avaliado por um procedimento de validação cruzada (*20 folds*). A avaliação também abordou um estudo de caso com uma lista de mutações encontradas em diferentes tipos de cânceres.

5.5 Comparativo entre os trabalhos relacionados

Finalizando este capítulo, a Tabela 14 resume as principais características apresentadas por cada trabalho relacionado a esta dissertação. A escolha de cada um deles partiu de artigos que avaliaram ferramentas de predição do impacto de mutações pontuais bem como por buscas realizadas em bases de dados científicas.

Embora a observação do uso potencial da combinação de algoritmos já tenha sido discutida na literatura, alguns dos trabalhos anteriores a esta dissertação tiveram como foco a avaliação e comparação do desempenho de diferentes modelos de predição do impacto de mutações pontuais com o objetivo principal de fornecer uma diretriz sobre o uso de métodos de AM para essas tarefas. Além disso, alguns trabalhos também apresentaram uma ferramenta baseada na metodologia proposta, logo foram abordados neste capítulo.

É importante destacar que o Capítulo 4 também abordou ferramentas análogas a aqui proposta, no entanto que fazem parte do *ensemble* proposto. As ferramentas apresentadas neste capítulo não puderam ser incluídas pelo fato de não cumprirem algum dos pré-requisitos de implementação estabelecidos.

Tabela 14: Comparativo entre os trabalhos relacionados: principais características.

Ferramenta	Principais características	Referência
iStable	Uma ferramenta integrada construída utilizando informações de sequências de proteínas e resultados de predição de diferentes preditores análogos. Vários métodos de AM foram avaliados na qual foi adotado o SVM. A ferramenta está disponível em dois tipos de entrada diferentes: estrutural e sequencial.	(CHEN; LIN; CHU, 2013)
Meta-tool (KStar)	A criação de uma meta-ferramenta que combina as saídas de sete ferramentas e um método adequado de consenso. As habilidades gerais de predição foram validadas em um conjunto de dados de teste composto de mutações de aminoácidos em múltiplos pontos. O algoritmo *KStar* obteve a mais alta acurácia de predição.	(MALINKA, 2015)
INPS	Uma ferramenta para predizer o efeito de mutações sobre a estabilidade de proteínas a partir de sua sequência. Foi mostrado que as predições do INPS são complementares às do mCSM (baseado em estrutura). Quando os dois métodos foram combinados, a acurácia sobre o conjunto *p53* foi maior do que a dos preditores individuais.	(FARISELLI et al., 2015)
ELASPIC	Usado para avaliar o efeito de mutações em proteínas do banco de dados *UniProt* e *Protein Data Bank*. É apoiado por um banco de dados com definições melhoradas de domínio estrutural bem como uma lista de interações domínio-domínio de proteínas. A avaliação da ferramenta apresentou também predições de uma lista de mutações causadoras de câncer.	(WITVLIET et al., 2016)

6 PROPOSTA EN-MUTATE

Este capítulo descreve a proposta intitulada EN-MUTATE e está dividido em duas seções principais: inicialmente é apresentada a metodologia em si; posteriormente, a ferramenta desenvolvida com base na proposta de *Ensemble Learning*. A Figura 29 apresenta o esquema de funcionamento do EN-MUTATE.

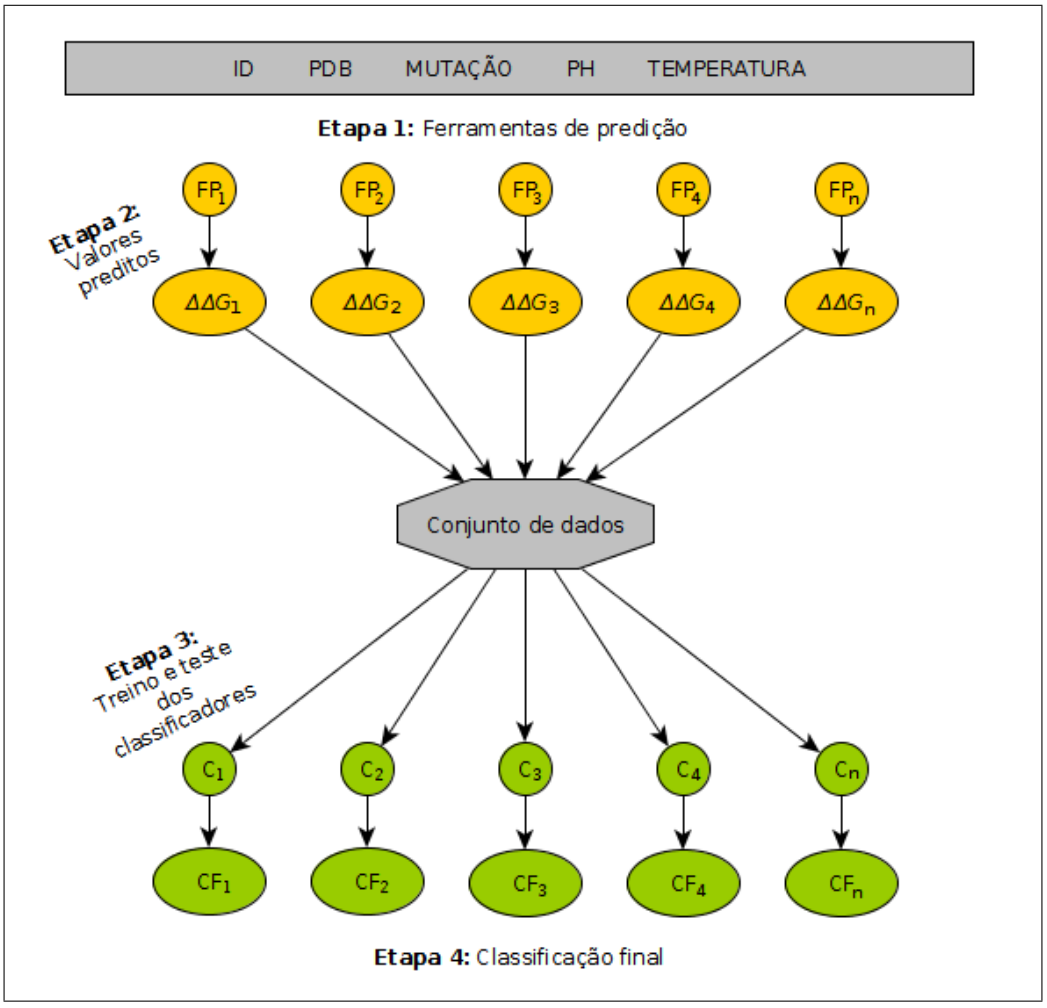


Figura 29: Esquema de funcionamento do EN-MUTATE.

6.1 Metodologia

A escolha de um bom método *a priori* pode ser decisória para o especialista obter um bom resultado. Com isso, o *Ensemble Learning* se torna uma alternativa à classificação baseada em um único método de aprendizado supervisionado. O EN-MUTATE consiste em uma abordagem que combina os resultados de predição de ferramentas correlatas, as quais servem de entrada para a criação de múltiplos modelos baseados em AM. Com isso, o objetivo é predizer qual o impacto que uma mutação pode resultar em um mutante *in-silico*. Sendo assim, a metodologia pode então ser usada para inferir mutações em uma proteína alvo. Cada uma das principais etapas que compõe o esquema de funcionamento do EN-MUTATE (Figura 29) é explicada nas próximas subseções.

6.1.1 Etapa 1: Ferramentas de predição

A predição do impacto de mutações pontuais através da metodologia EN-MUTATE parte da combinação de resultados de diferentes ferramentas de predição (FP_1, FP_2, FP_3, FP_4, FP_n - Figura 29) que recebem como entrada: a proteína (formato PDB), a mutação, a temperatura e o pH do experimento. A descrição dos valores de entrada é apresentada na Tabela 15.

Tabela 15: Descrição dos valores de entrada da metodologia EN-MUTATE

Atributo	Descrição	Exemplo
Id	Identificador da mutação dentro do conjunto de dados	1
PDB	Código PDB da proteína	1A23
Mutação	Aminácido nativo seguido da posição a ser alterada e o mutante	H32L
pH	pH do experimento	7
Temperatura	Temperatura do experimento	25

Os parâmetros exibidos na Tabela 15 bem como o valor de $\Delta\Delta G$ experimental das mutações pontuais - utilizado para a validação dos modelos preditivos - podem ser obtidos através de bancos de dados biológicos, como o *ProTherm* (BAVA et al., 2004). Desta forma, cada ferramenta de predição integrada é capaz de predizer um valor de $\Delta\Delta G$ da mutação informada, o que é abordado na Etapa 2 da metodologia proposta.

6.1.2 Etapa 2: Valores preditos

Uma vez que as tarefas de AM tendem a aprender mais com um número maior de instâncias disponíveis nos conjuntos de dados de treinamento (TAN et al., 2006), o uso de rotinas sistematizadas (*scripts*) na execução dos experimentos possibilita ao especialista uma melhor precisão na manipulação dos dados. no contexto deste trabalho, estudos

de casos compostos por um número maior de mutações pontuais. Os valores de saída preditos ($\Delta\Delta G_1$, $\Delta\Delta G_2$, $\Delta\Delta G_3$, $\Delta\Delta G_4$, $\Delta\Delta G_n$ - Figura 29) compõem o conjunto de dados que servirá para o treinamento dos modelos de AM. A Figura 30 ilustra um exemplo de arquivo no formato CSV (*Comma-separated values*) com dez instâncias, na qual os atributos ($\Delta\Delta G$) e a classe (*Experimental_Classification*) são separados por vírgulas.

	M	O	Q	S	U	AE
1	ΔΔG1	ΔΔG2	ΔΔG3	ΔΔG4	ΔΔGn	Experimental_Classification
2	1.05	0.08	2.24	0.80	0.81	Stabilizing
3	-1.06	-2.18	5.72	-1.44	-1.20	Destabilizing
4	-0.33	0.68	-3.22	-2.03	-2.21	Neutral
5	-0.56	-5.27	-3.69	-1.65	-1.75	Neutral
6	0.48	1.66	-1.72	-0.13	-0.17	Destabilizing
7	0.61	0.66	0.6	1.29	1.27	Stabilizing
8	-1.57	-18.04	-0.55	-1.06	-1.14	Neutral
9	-0.46	0.49	1.12	-1.43	-1.07	Destabilizing
10	-1.03	-4.41	-3.94	-1.37	-1.57	Destabilizing

Figura 30: Arquivo CSV com os valores de saída das ferramentas de predição.

Após a definição do conjunto de dados (Figura 30) que servirá de entrada para as tarefas de classificação, a Etapa 3 aborda o treinamento e, posteriormente, o teste dos classificadores baseados em AM.

6.1.3 Etapa 3: Treino e teste dos classificadores

Esta etapa se refere a classificação pelos algoritmos de AM (C_1, C_2, C_3, C_4, C_n - Figura 29). A literatura descreve diferentes técnicas de aprendizado supervisionado que exploram particularidades distintas (WITTEN; FRANK; HALL, 2011; HASTIE; TIBSHIRANI; FRIEDMAN, 2011), sendo possível integrá-las ao EN-MUTATE. Também nesta etapa, o atributo classe ($\Delta\Delta G$ experimental) é discretizado em intervalos referentes a estabilidade de proteínas, tendo como rótulos: desestabilizante, neutra ou estabilizante.

A Figura 31 apresenta a interface de classificação (*Classify*) do WEKA[1] (*Waikato Environment for Knowledge Analysis*), uma plataforma escrita em JAVA contendo um conjunto de algoritmos de AM e ferramentas de pré-processamento de dados (WITTEN; FRANK; HALL, 2011). Através do WEKA, por exemplo, é possível treinar e testar os modelos preditivos que farão parte do EN-MUTATE.

Durante o treinamento é utilizado o conjunto de dados exemplificado na Figura 30 contendo os valores de saída ($\Delta\Delta$) das ferramentas de predição bem como o $\Delta\Delta G$ experimental discretizado. Com isso, são definidas as abordagens para construir os modelos de classificação a partir desses dados de entrada são aplicados algoritmos de AM para induzir o modelo que melhor identifica as relações entre esses atributos (HAN; PEI; KAMBER, 2006). A validação (teste) do modelo pode ser realizada com os próprios dados de treinamento ou através de diretivas que reservam uma certa quantidade de instâncias para

[1] http://www.cs.waikato.ac.nz/ml/weka/downloading.html

treinamento e utiliza o restante para o teste (WITTEN; FRANK; HALL, 2011). Dessa forma, as taxas de erro nas diferentes iterações são calculadas para produzir uma taxa de erro global (HASTIE; TIBSHIRANI; FRIEDMAN, 2011).

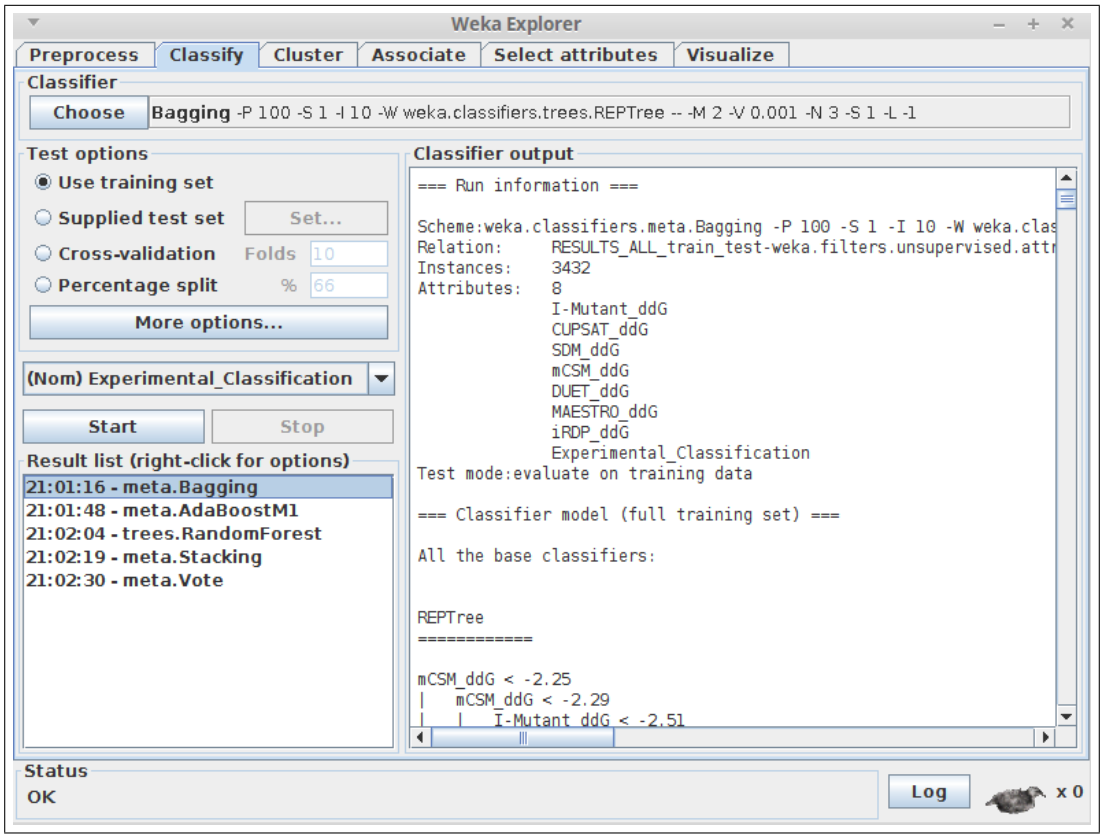


Figura 31: Interface de classificação do WEKA.

A próximo passo da metodologia EN-MUTATE é a classificação final, ou seja, a predição de novas instâncias a ser realizada pelos modelos definidos.

6.1.4 Etapa 4: Classificação final

Após a definição dos modelos preditivos (CF_1, CF_2, CF_3, CF_4, CF_n - Figura 29), finalmente a metodologia EN-MUTATE é capaz de predizer novas instâncias a ela submetidas. Tais predições podem ser realizadas individualmente por cada um dos modelos, ou, ainda, de maneira unificada através de consenso. A Figura 32 mostra um exemplo de um classificador baseado em árvore de decisão obtido a partir de diferentes ferramentas de predição do impacto de mutações em proteínas (FP_1, FP_2, FP_3). Nesse exemplo ilustrado é apresentada uma árvore de decisão para indicar qual é a classe (Desestabilizante, Neutra ou Estabilizante) referente ao impacto de uma mutação pontual em proteína dado um experimento de substituição de um aminoácido no gene baseado no valor de $\Delta\Delta G$ gerado por diferentes ferramentas computacionais.

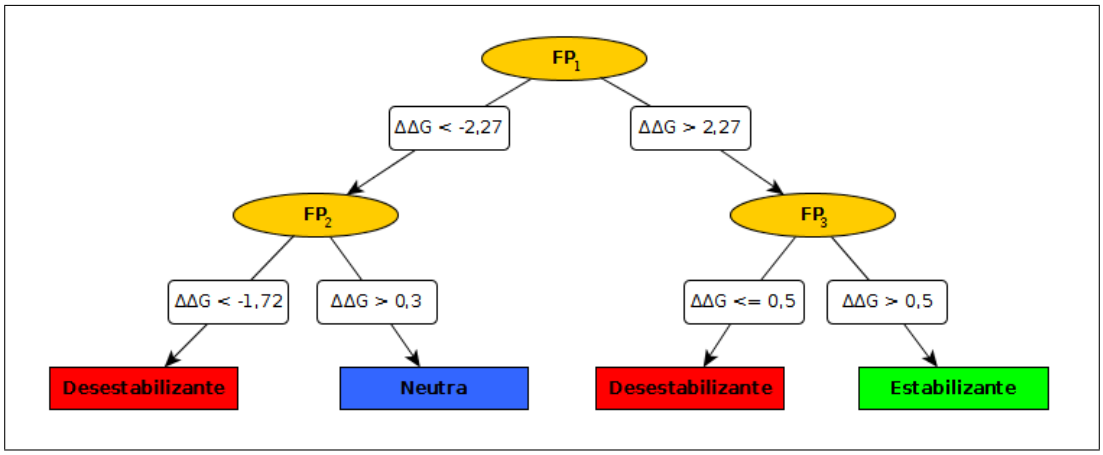


Figura 32: Exemplo de um modelo preditivo baseado em árvore de decisão.
Fonte: Adaptado de (WITTEN; FRANK; HALL, 2011).

Percorrendo a árvore de decisão da Figura 32 da raiz às folhas é possível verificar, por exemplo, que se o valor indicado pela ferramenta mCSM for maior do que 2,27 Kcal/mol e o valor de saída da ferramenta SDM for maior que 0,5 Kcal/mol, a mutação irá estabilizar a proteína. Os atributos preditivos são os valores de $\Delta\Delta G$ uma vez que o atributo classe é o impacto da mutação.

6.2 Ferramenta

O método baseado em *Ensemble Learning* aqui proposto resultou na implementação de uma ferramenta denominada EN-MUTATE*web*, criada para permitir que a predição de mutações pontuais através da metodologia EN-MUTATE possa ser realizada de uma maneira mais amigável aos usuários. Assim sendo, pesquisadores de diferentes áreas poderão ter acesso à plataforma que será disponibilizada via *web*, tendo o seu acesso realizado diretamente pelo navegador. O esquema de predição que embasa a ferramenta está ilustrado na Figura 33.

O uso do EN-MUTATE*web* parte da submissão da estrutura da proteína ou da identificação do seu código, ambos no formato PDB (*Protein Data Bank*), seguidos da especificação da mutação a ser predita bem como os parâmetros do experimento. O fluxo de execução da ferramenta é então iniciado pelos valores de entrada definidos pelo usuário.

A etapa subsequente de execução do EN-MUTATE*web* refere-se, exclusivamente, à designação das tarefas para as ferramentas de predição integradas bem como a manipulação dos resultados, ambas as tarefas executadas em *background* no servidor de aplicação. Ao final das predições, os resultados são exibidos ao usuário. A proposta inicial compreendeu um *ensemble* composto por sete ferramentas (*I-Mutant*, CUPSAT, SDM, mCSM, DUET, iRDP e MAESTRO (LAIMER et al., 2015)). No entanto, a sua implementação permite a inclusão de novas ferramentas sem que o seu núcleo seja alterado.

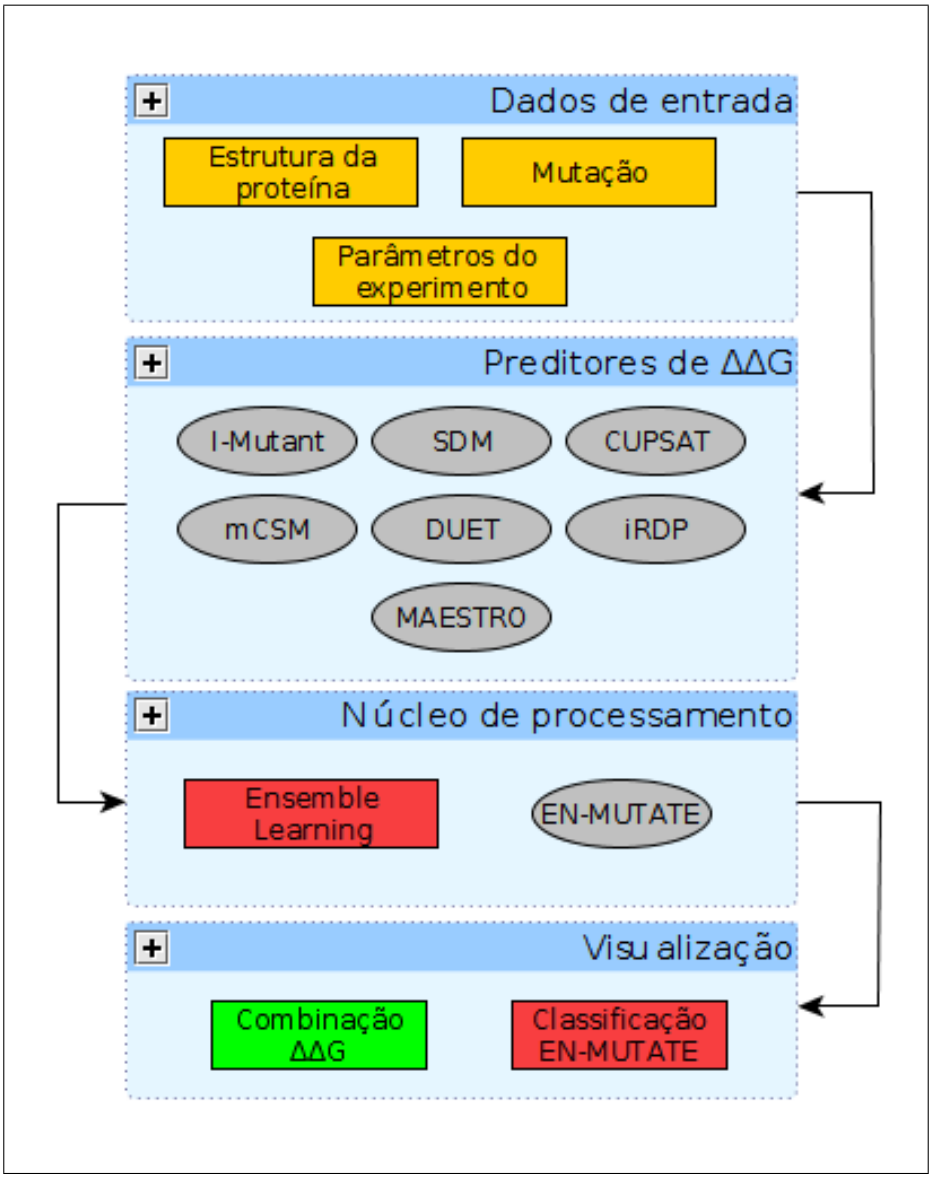


Figura 33: Esquema de funcionamento do EN-MUTATE*web*.

A Tabela 16 reúne as principais tecnologias adotadas no desenvolvimento do EN-MUTATE*web*.

Tabela 16: EN-MUTATE*web*: características de implementação.

Tecnologia	Principal funcionalidade	Documentação
PHP	Processamento *backend*	http://php.net
HTML	Estrutura das páginas *web*	http://w3schools.com/html
CSS	Formatação das páginas *web*	http://w3schools.com/css
jQuery	Manipulação *frontend*	http://jquery.com
SQL	Consulta ao banco de dados	http://sql.org
Bootstrap	Adaptação para dispositivos móveis	http://getbootstrap.com
Python	Manipulação de arquivos	http://python.org
iMacros	Submissão as ferramentas integradas	http://imacros.net
MySQL	Sistema Gerenciador de Banco de Dados	http://mysql.com
WEKA	Criação e instanciação dos modelos	http://waikato.ac.nz/ml/weka

A linguagem de programação PHP foi escolhida pela sua concepção voltada para o desenvolvimento de ferramentas *web*, possibilitando, dessa maneira, a integração com as demais tecnologias empregadas. O uso do CSS combinado ao HTML permitiu a padronização da interface do EN-MUTATE*web*, sendo as principais linguagens *frontend*. Da mesma maneira, o *JavaScript* (*framework jQuery*) foi utilizado para a validação das entradas dos formulários, além da manipulação dos menus bem como dos efeitos de transição. O *framework Bootstrap* também foi adotado para construir a interface da ferramenta. Sua escolha se deu por permitir a adaptação da interface em diferentes tamanhos de tela, incluindo dispositivos móveis, como *tablets* e *smartphones*. Para as tarefas *backend*, através do uso da linguagem de programação *Python* foi possível a execução sistemática de importantes rotinas da proposta, desse modo, foi atribuída como a linguagem de manipulação do núcleo do *EN-MUTATEweb*.

Grande parte dos procedimentos para a geração dos arquivos no formato padrão aceito por cada ferramenta de predição adotada, assim como a manipulação desses valores, foram implementados através de *scripts* de automatização. Para as tarefas de *web scraping* - extração de dados de *websites* - o *framework iMacros* foi adotado já que possui uma linguagem de programação própria, permitindo que rotinas sejam automatizadas, por exemplo, simulando uma navegação manual na *web*. Dessa forma, no *EN-MUTATEweb*, o acesso as ferramentas de predição é feito integralmente pelo *framework iMacros*. O *MySql* e a linguagem SQL foram escolhidos para armazenar e organizar os dados processados. Por fim, durante a etapa de preparação dos dados, são gerados os arquivos ARFF (*Attribute-Relation File Format*) para serem utilizados na etapa de mineração de dados pela plataforma WEKA.

6.2.1 Banco de dados

Um dos requisitos de desenvolvimento do *EN-MUTATEweb* foi a definição da sua base de dados. O projeto buscou permitir a inclusão dinâmica de novos preditores ao *ensemble*, ou seja, sem a necessidade de atualização da estrutura das tabelas. A descrição da base de dados do *EN-MUTATEweb* é exibida na Figura 34, sendo representada por um diagrama que mostra graficamente os relacionamentos entre as tabelas.

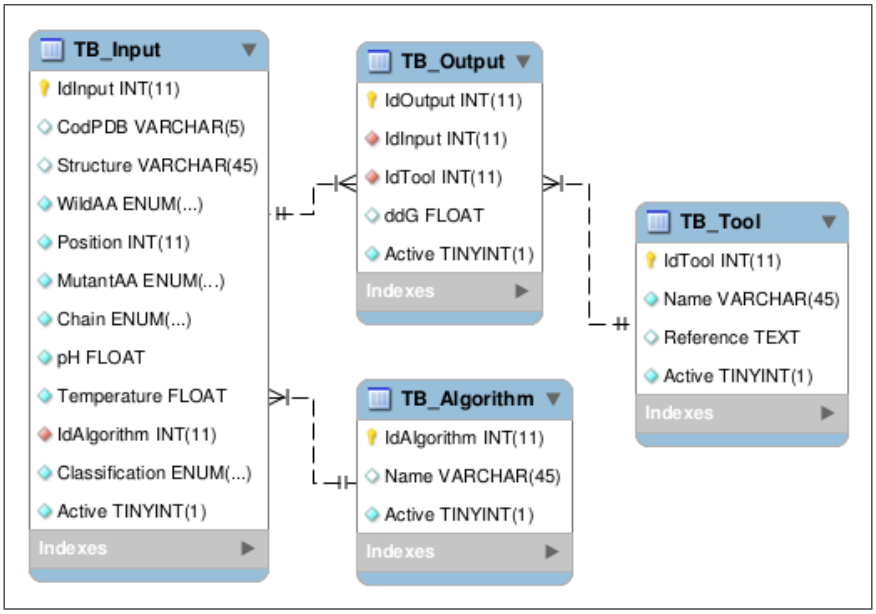


Figura 34: Detalhes das tabelas que compõem o EN-MUTATE*web*. A nomenclatura está em Inglês.

A principal tabela em que são armazenados os valores referentes a submissão inicial feita pelo usuário foi nomeada "*TB_Input*". As colunas "*CodPDB*", "*Structure*" e "*Chain*" armazenam as informações pertences à estrutura submetida. As colunas "*WildAA*", "*Position*" e "*MutantAA*" compõem os dados referentes a mutação pontual definida pelo usuário. As colunas "*pH*" e "*Temperature*" armazenam os dados experimentais a serem utilizados como parâmetros da predição, seguidos da coluna "*Classification*", na qual identifica a classificação binária (Desestabilizante, Estabilizante) ou ternária (Desestabilizante, Neutra, Estabilizante) dos resultados. As tabelas "*TB_Algorithm*" e "*TB_Tool*" armazenam os nomes dos algoritmos bem como das ferramentas de predição que integram a metodologia EN-MUTATE, por isso possuem uma semelhança entre si.

A tabela "*TB_Output*" representa os dados de saída (coluna "*ddG*") dos preditores, estando relacionada as tabelas "*TB_Input*" e "*TB_Tool*". Por questões de segurança, as chaves estrangeiras foram configuradas com a opção "*Restrict*", evitando que a alteração de algum registro possa afetar a integridade da base de dados. Por fim, um detalhe a ser mencionado é que as tabelas já foram projetadas para receber, dentro do possível, valores pré-estabelecidos em uma lista (tipo de dado ENUM), somente números (tipos de dados INT ou FLOAT) assim como *strings* limitadas (tipo de dado VARCHAR).

6.2.2 Interface

Prosseguindo com a descrição da interface da ferramenta, a Tabela 17 reúne os valores aceitos como entrada no EN-MUTATE*web*. É importante ser destacado que todos os campos possuem dicas de uso (*tooltips*) e um processo de validação realizado em 3 etapas: pelo navegador (*JavaScript*), pelo servidor de aplicação (PHP) e pelo sistema gerenciador de banco de dados (*MySQL*).

Tabela 17: Campos de entrada do EN-MUTATE*web*.

Rótulo	Valores aceitos	
Specify structure	Código PDB (4 caracteres alfanuméricos) ou a estrutura da proteína (até 5MB)	*
Chain	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z	*
Amino acid (native)	ALA (A), CYS (C), ASP (D), GLU (E), PHE (F), GLY (G), HIS (H), ILE (I), LYS (K), LEU (L), MET (M), ASN (N), PRO (P), GLN (Q), ARG (R), SER (S), THR (T), VAL (V), TRP (W), TYR (Y)	*
Position	Valores inteiros positivos	*
Amino acid (mutant)	ALA (A), CYS (C), ASP (D), GLU (E), PHE (F), GLY (G), HIS (H), ILE (I), LYS (K), LEU (L), MET (M), ASN (N), PRO (P), GLN (Q), ARG (R), SER (S), THR (T), VAL (V), TRP (W), TYR (Y)	*
pH	Valores reais positivos ou negativos	
Temperature	Valores reais positivos ou negativos	
Ensemble algorithm	BAGGING, BOOSTING, RANDOM FORESTS, STAKING, VOTING	*
$\Delta\Delta G$ *classification*	BINARY (DESTABILIZING, STABILIZING), TERNARY (DESTABILIZING, NEUTRAL, STABILIZING)	*

* Campos obrigatórios

Para demonstrar a operacionalidade da ferramenta desenvolvida, esta seção também apresenta os detalhes da interface do EN-MUTATE*web*. A Figura 35 exibe a sua tela inicial. As opções de navegação incluem 3 possibilidades:

- ***Submit***: o usuário será direcionado ao formulário de submissão de uma nova tarefa (denominada "*Job*");
- ***About***: são exibidas as informações sobre o esquema de predição adotado pela ferramenta;
- ***Contact***: é apresentado um formulário de contato com o ComBi-Lab (*Computional Biology Laboratory*) da FURG (Universidade Federal do Rio Grande), responsável direto pela administração da ferramenta.

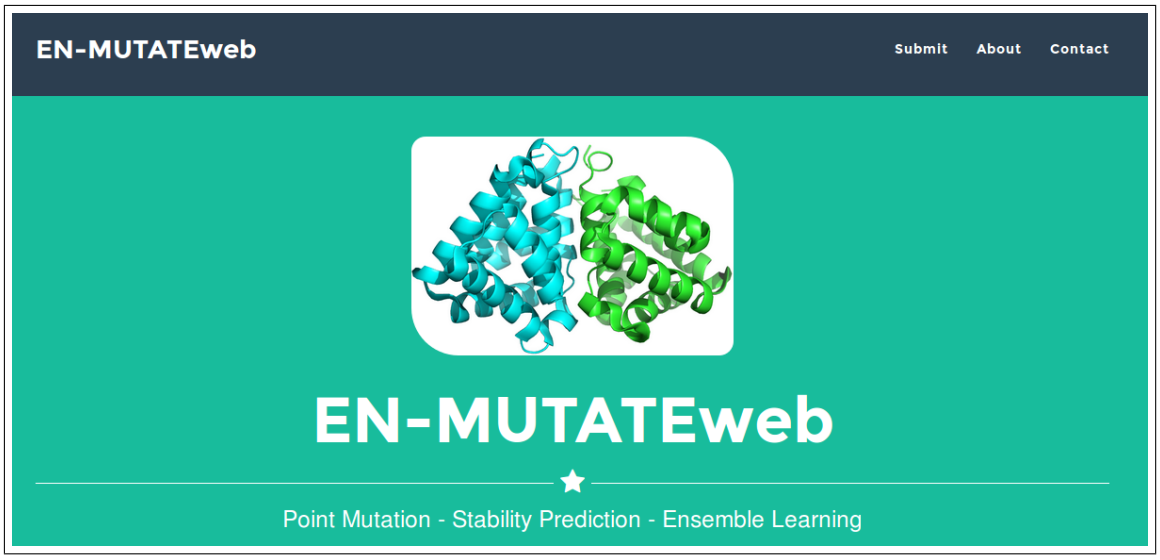


Figura 35: Funcionamento do EN-MUTATE*web*: apresentação.

O formato de submissão da estrutura da proteína inclui um arquivo com as coordenadas tridimensionais dos átomos que constituem a molécula no padrão definido pelo PDB. A especificação da mutação é dada pela combinação do aminoácido nativo, seguido da posição a ser alterada e o novo aminoácido.

No exemplo exibido na Figura 36, para realizar a predição do impacto de uma mutação pontual na proteína de código PDB: 1A23 (opção "*Specify structure*"), cadeia A (*Chain*), foi realizada a substituição do aminoácido histidina (opção "*Amino acid (native)*") por uma leucina (opção "*Amino acid (mutant)*") na posição 32 (*Position*), com os parâmetros de pH e temperatura (*Temperature*) preenchidos com os seus valores padrão, 7 e 25, respectivamente.

Para essa demonstração foi selecionado o algoritmo *Stacking* (opção "*Ensemble algorithm*"), seguido de uma classificação ternária de $\Delta\Delta G$ (opção "$\Delta\Delta G$ *classification*"). Ao clicar no botão "*Submit to EN-MUTATEweb*"uma nova tarefa (*job*) é criada. A ferramenta também possibilita a execução de um exemplo (botão "*Run example*") no qual os campos são preenchidos automaticamente.

Caso o preenchimento dos dados de entrada estiverem corretos, o formulário de submissão da ferramenta direciona o usuário para a página de exibição dos resultados, detalhada na Figura 37. Cada nova tarefa gera um identificador único (JOB ID) que fica armazena no banco de dados, diante disso, o acesso pode ser realizado diretamente via URL, mesmo após o fechamento do navegador de *Internet*.

A página de resultados é atualizada a cada 30 segundos até que a tarefa seja executada por completo. Os resultados são exibidos em 3 tabelas: a primeira e a segunda ("*Mutation site*" e "*Prediction parameters*") compõem os dados de entrada submetidos pelo usuário, seguidos da terceira tabela ("*Predicted stability change*"), na qual são preenchidos os valores preditos por cada ferramenta integrante do *ensemble* adotado.

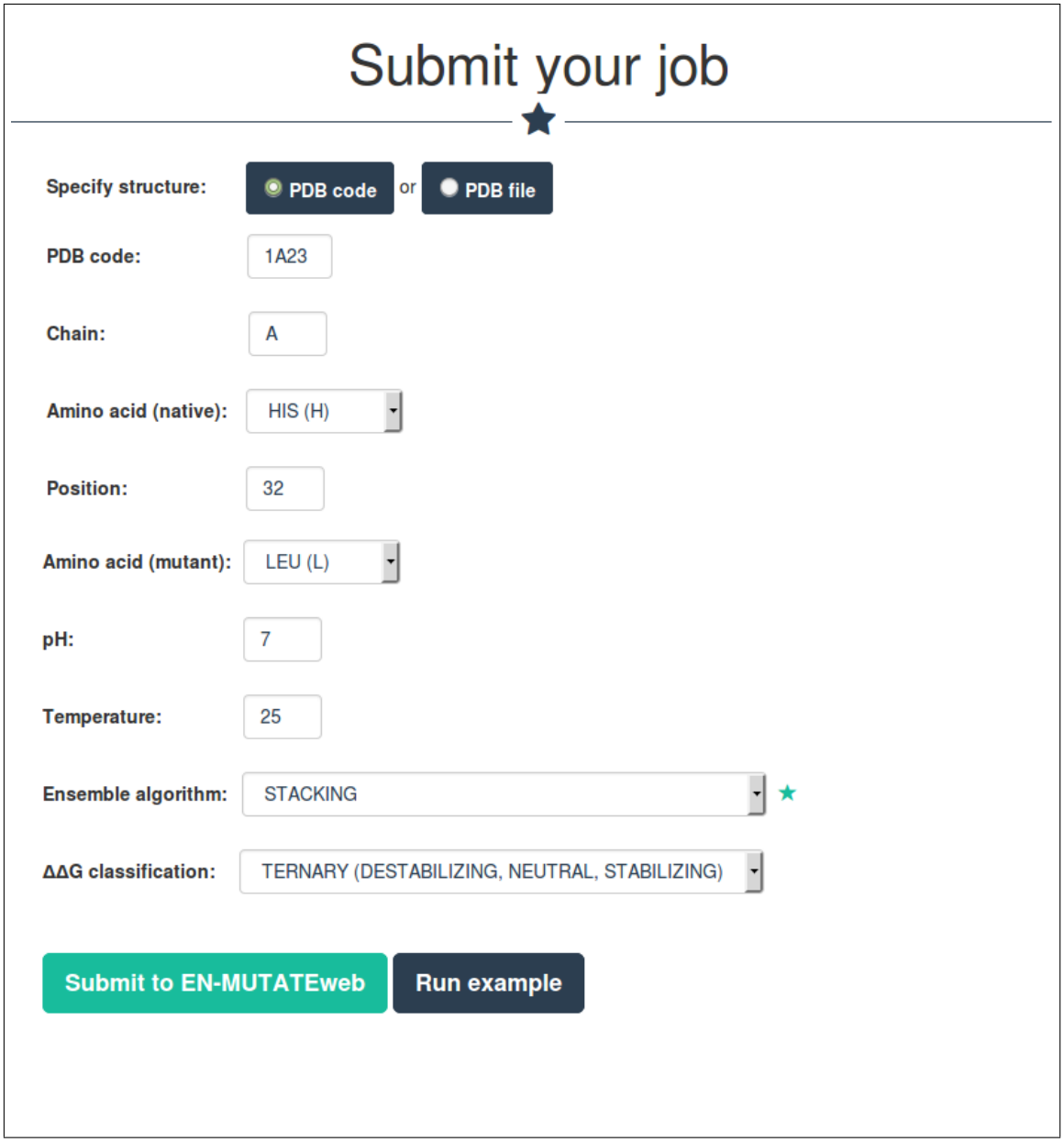
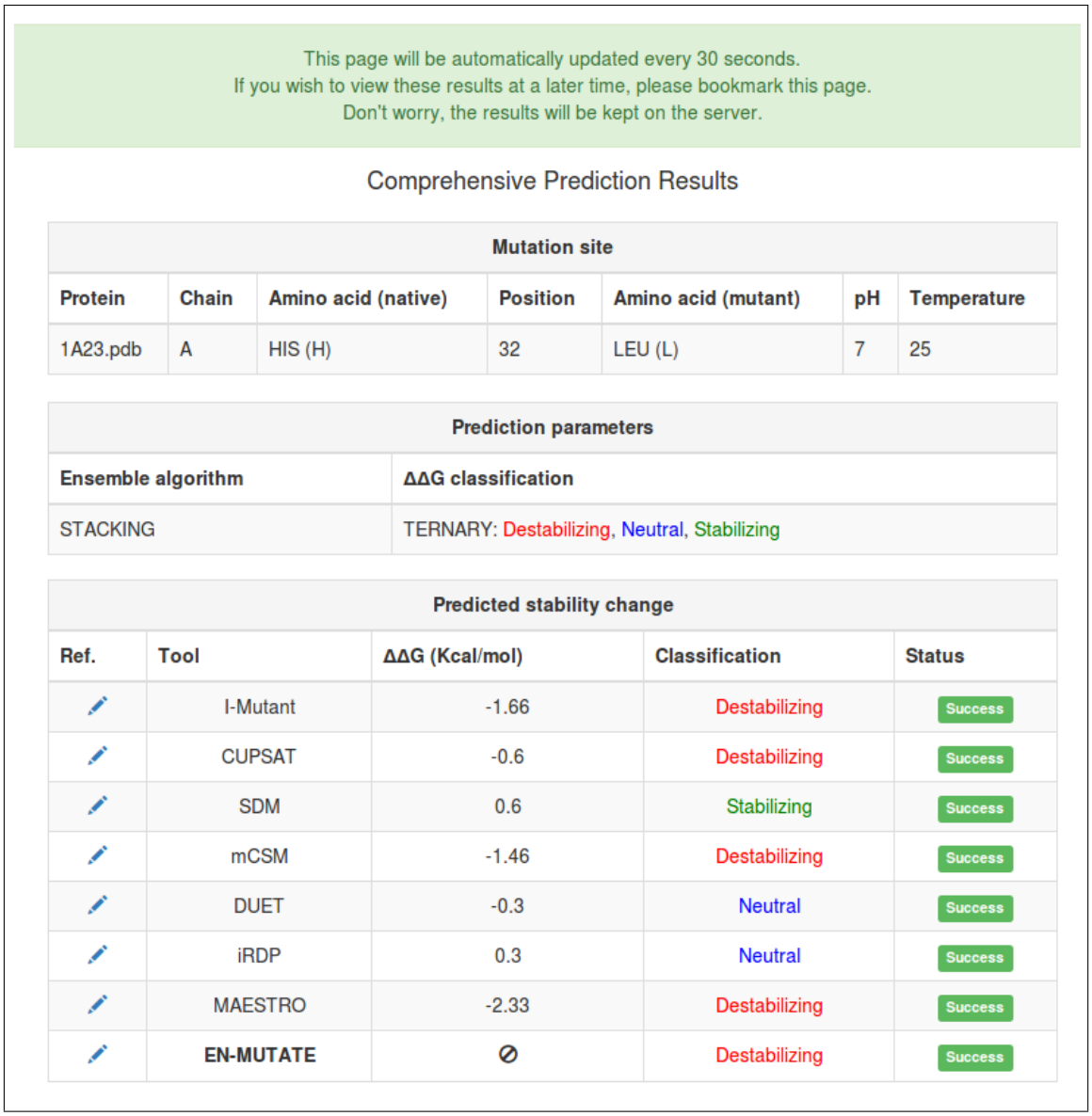


Figura 36: Exemplo de uso (submissão): EN-MUTATE*web*.



This page will be automatically updated every 30 seconds.
If you wish to view these results at a later time, please bookmark this page.
Don't worry, the results will be kept on the server.

Comprehensive Prediction Results

Mutation site						
Protein	**Chain**	**Amino acid (native)**	**Position**	**Amino acid (mutant)**	**pH**	**Temperature**
1A23.pdb	A	HIS (H)	32	LEU (L)	7	25

Prediction parameters	
Ensemble algorithm	**ΔΔG classification**
STACKING	TERNARY: Destabilizing, Neutral, Stabilizing

Predicted stability change				
Ref.	**Tool**	**ΔΔG (Kcal/mol)**	**Classification**	**Status**
	I-Mutant	-1.66	Destabilizing	Success
	CUPSAT	-0.6	Destabilizing	Success
	SDM	0.6	Stabilizing	Success
	mCSM	-1.46	Destabilizing	Success
	DUET	-0.3	Neutral	Success
	iRDP	0.3	Neutral	Success
	MAESTRO	-2.33	Destabilizing	Success
	EN-MUTATE	⊘	Destabilizing	Success

Figura 37: Exemplo de uso (saída): EN-MUTATE*web*.

7 RESULTADOS E DISCUSSÃO

Este capítulo descreve os resultados obtidos em diferentes experimentos através da aplicação de *Ensemble Learning* pelo método proposto denominado EN-MUTATE. As análises aqui abordadas foram baseadas nos conjuntos descritos na seção 7.2 e os experimentos foram divididos em cinco abordagens. Na primeira delas foi aplicada uma votação por pluralidade diretamente aos valores discretizados preditos pelas ferramentas integrantes do *ensemble*. Nas demais abordagens, os modelos preditivos foram avaliados utilizando tanto o mesmo conjunto para treinamento e teste quanto através de uma escolha aleatória de 70% dos dados para treinamento e 30% para teste, sendo que a discretização do valor de $\Delta\Delta G$ foi realizada em duas e três classes, conforme a descrição abordada no Capítulo 2. Por ser mais robusta ao desbalanceamento de classes, a Medida-F (micro média) - ponderada pelo número de instâncias de cada classe - foi adotada como principal métrica de avaliação dos experimentos. Também é importante destacar que os valores de precisão e revocação apresentados estão considerando a média aritmética entre as classes que compõem os conjuntos de dados utilizados.

7.1 Descrição dos algoritmos

Os experimentos de *Ensemble Learning* desta dissertação foram realizados através da plataforma WEKA (WITTEN; FRANK; HALL, 2011) e as demais análises foram implementadas com a linguagem de programação *Python* através da *scikit-learn*[1], um conjunto de bibliotecas específicas para AM. Os resultados foram gerados utilizando os parâmetros padrão dos algoritmos de AM bem como de cada ferramenta de predição comparada. A Tabela 18 resume as principais características apresentadas por cada classificador aplicado a metodologia EN-MUTATE. Os parâmetros de execução são listados na Tabela 19.

[1] http://scikit-learn.org

Tabela 18: Comparativo entre os classificadores adotados: principais características.

Classificador	Principais características	Referência
J48	Uma implementação na linguagem JAVA do algoritmo baseado em árvore de decisão C4.5 escrito por Quinlan (1993).	(WITTEN; FRANK; HALL, 2011)
SVM	Uma combinação linear de padrões de suporte nas quais são o subconjunto de padrões de treinamento que estão mais próximos do limite de decisão, sendo aplicável a uma variedade de funções de classificação.	(BOSER; GUYON; VAPNIK, 1992)
MultilayerPerceptron	Uma arquitetura de rede neural bastante variável, mas que em geral consiste de várias camadas de neurônios. A camada de entrada introduz o vetor inicial para a rede que pode ter uma ou mais camadas ocultas e uma camada de saída.	(GARDNER; DORLING, 1998)
NaiveBayes	Abordagens probabilísticas que fazem fortes suposições sobre como os dados são gerados e postulam um modelo probabilístico que incorpora essas premissas.	(JOHN; LANGLEY, 1995)
RandomForest	Uma combinação de preditores de árvore em que cada árvore depende dos valores de um vetor aleatório *bagging* amostrado independentemente e com a mesma distribuição para toda a "floresta".	(BREIMAN, 2001)
Bagging	Define amostras separadas do conjunto de dados de treinamento bem como um classificador para cada amostra.	(BREIMAN, 1996)
AdaBoostM1	Em vez de utilizar uma sucessão de amostras de inicialização independentes como o *Bagging*, atribui um peso para individual para as instâncias e a cada iteração um vetor de pesos é ajustado.	(FREUND; SCHAPIRE et al., 1996)
Stacking	Múltiplos algoritmos distintos são aplicados aos dados de treinamento e posteriormente um meta-classificador combinador é treinado para realizar a predição final.	(DZEROSKI; ZENKO, 2004)
Cascading	Pode ser considerado como um caso especial de *Stacking*, no entanto, as predições são adicionadas ao conjunto de dados inicial e passadas para o próximo nível.	(GAMA; BRAZDIL, 2000)
Vote	Serve para combinar classificadores. Existem diferentes combinações de estimativas de probabilidade para classificação, como a votação majoritária, média ou mediana.	(KITTLER et al., 1998)

Tabela 19: Comparativo entre os classificadores adotados: parâmetros de execução.

Classificador	Parâmetros de execução
J48	*binarySplits: False; confidenceFactor: 0,25; debug: False; minNumObj: 2; numFolds: 3; reducedErrorPruning: False; saveInstanceData: False; seed: 1; subtreeRaising: True; unpruned: False; useLaplace: False.*
SVM	*SVMType: C-SVC (Classification); cacheSize: 40; coef0: 0;cost: 1; debug: False; degree: 3; doNotReplaceMissingValues: False; eps: 0,001; gamma: 0; kernelType: RBF; loss: 0,1; normalize: False; nu: 0,5; probabilityEstimates: False; seed: 1; shrinking: True; weights: 0.*
MultilayerPerceptron	*GUI: False; autoBuild: True; debug: False; decay: False; hiddenLayers: a; learningRate: 0,3; momentum: 0,2; nominalToBinaryFilter: True; normalizeAttributes: True; normalizeNumericClass: True; reset: True; seed: 0; trainingTime: 500; validationSetSize: 0; validationThreshold: 20.*
NaiveBayes	*debug: False; displayModelInOldFormat: False; useKernelEstimator: False; useSupervisedDiscretization: False.*
RandomForest	*debug: False; maxDepth: 0; numFeatures: 0; numTrees: 10; seed: 1.*
Bagging	*bagSizePercent: 100; calcOutOfBag: False; classifier: [J48, SVM, MultilayerPerceptron, NaiveBayes]; debug: False; numIterations: 10; seed: 1.*
AdaBoostM1	*classifier: [J48, SVM, RandomForest, MultilayerPerceptron, NaiveBayes]; debug: False; numIterations: 10; seed: 1; useResampling: False; weightThreshold: 100.*
Stacking	*classifiers: [J48, RandomForest, MultilayerPerceptron, NaiveBayes]; debug: False; metaClassifier: SVM; numFolds: 10; seed: 1.*
Vote	*classifiers: [J48, SVM, RandomForest, MultilayerPerceptron, NaiveBayes]; combinationRule: Majority voting; debug: False seed: 1.*

7.2 Descrição dos conjuntos de dados

Os conjuntos de dados de entrada utilizados nesta dissertação partiram da seleção de resultados experimentais provenientes dos bancos de dados biológicos *ProTherm* (dados termodinâmicos de mutações) e *Protein Data Bank* (estrutura das proteínas). Esses valores serviram como parâmetros para a predição realizada pelas ferramentas integrantes do *ensemble* do mesmo modo para a avaliação dos modelos preditivos. A Figura 38 ilustra o esquema de distribuição dos conjuntos de dados utilizados nos experimentos.

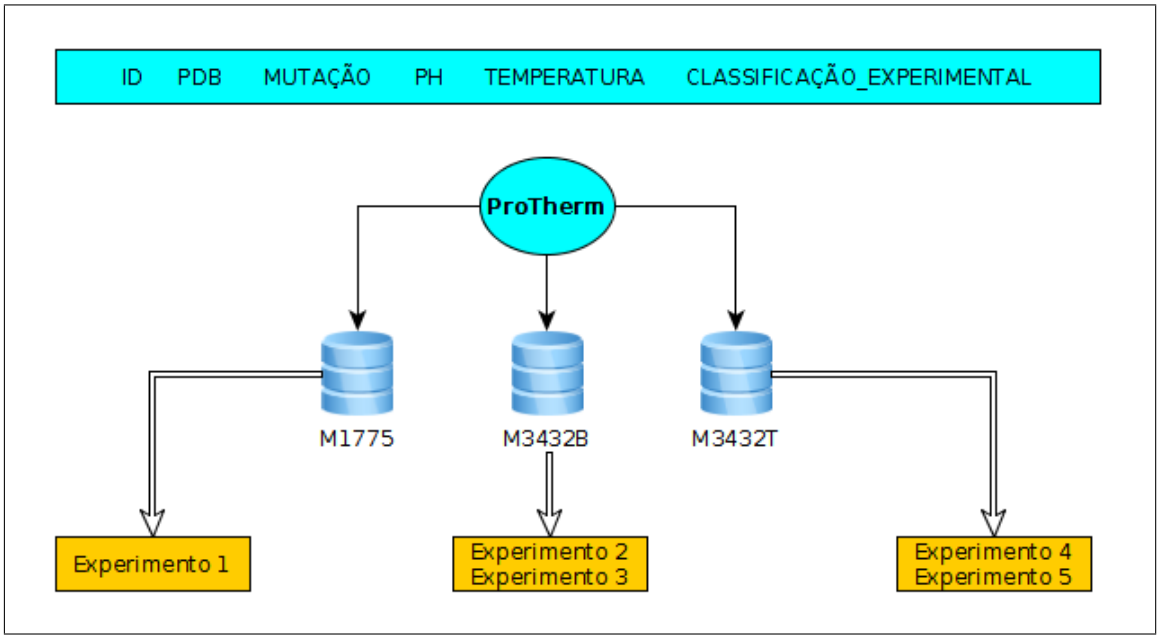


Figura 38: Conjuntos de dados utilizados nos experimentos: esquema de distribuição.

A descrição dos conjuntos de dados, denominados como M1775 (HUANG; GROMIHA; HO, 2007) e M3432B/M3432T (nesse caso, "B" com classes binárias e "T" com classes ternárias), é apresentada na Tabela 20 e complementada na Tabela 21. As mutações pontuais estão distribuídas em 179 proteínas.

Tabela 20: Conjuntos de dados utilizados: características gerais.

Nome	Instâncias	Atributos	Classificação
M1775	1775	6	Desestabilizante (946), Neutra (552), Estabilizante (277)
M3432B	3432	6	Desestabilizante (2217), Estabilizante (1215)
M3432T	3432	6	Desestabilizante (1885), Neutra (940), Estabilizante (607)

Tabela 21: Conjuntos de dados utilizados: descrição dos atributos.

Atributo	Descrição	Exemplo
Id	Identificador da mutação dentro do conjunto de dados	777
PDB	Código PDB da proteína	1CAH
Mutação	Aminácido nativo seguido da posição a ser alterada e o mutante	S56C
pH	pH do experimento	7,5
Temperatura	Temperatura do experimento	23
Classificação_experimental	Classificação do resultado	Desestabilizante

7.3 Experimento 1: conjunto de dados M1775 utilizando votação por pluralidade

O primeiro experimento a ser apresentado neste capítulo utilizou um *ensemble* através da técnica de votação por pluralidade (*plurality voting*) baseado na predição das sete ferramentas integradas: *I-Mutant*, CUPSAT, SDM, mCSM, DUET, iRDP, MAESTRO. É importante destacar que esse experimento teve como objetivo aferir a viabilidade do uso de *Ensemble Learning* para a predição do impacto de mutações pontuais em proteínas e serviu como base para a definição dos demais. A classe ($\Delta\Delta G$) foi discretizada em três categorias: desestabilizante (946 instâncias), neutra (552 instâncias) e estabilizante (277 instâncias). Os melhores resultados comparados obtidos com o conjunto de dados simplificado denominado M1775 (1775 instâncias) estão destacados em negrito na Tabela 22 e posteriormente ilustrados na Figura 39 através de suas respectivas matrizes de confusão. Em matrizes de confusão, a diagonal principal representa os valores preditos corretamente.

Tabela 22: Experimento 1: conjunto M1775 utilizando votação por pluralidade.

Ensemble de classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
Plurality voting	**73,13%**	**0,740**	**0,731**	**0,721**
Ferramentas				
Preditor	Acurácia	Precisão	Revocação	Medida-F
I-Mutant	**76,06%**	**0,763**	**0,761**	**0,753**
CUPSAT	65,63%	0,656	0,656	0,656
SDM	50,42%	0,556	0,504	0,517
mCSM	59,94%	0,580	0,599	0,577
DUET	59,55%	0,616	0,595	0,588
iRDP	56,79%	0,609	0,568	0,582
MAESTRO	61,58%	0,632	0,616	0,617

EN-MUTATE: Plurality voting

=== Matriz de Confusão ===

a	b	c	<-- classificado como
99	50	128	a = Estabilizante
4	819	123	b = Desestabilizante
20	152	380	c = Neutra

(a)

Ferramenta: I-Mutant

=== Matriz de Confusão ===

a	b	c	<-- classificado como
106	24	147	a = Estabilizante
30	821	95	b = Desestabilizante
18	111	423	c = Neutra

(b)

Figura 39: Experimento 1: Matrizes de confusão dos preditores com melhores resultados.

Os resultados do total das 1775 predições avaliadas neste experimento tiveram como melhores preditores, considerando a Medida-F, o *I-Mutant* (0,753) seguido do método aqui proposto EN-MUTATE: *Plurality voting* (0,721). Em contrapartida, as ferramentas SDM (0,517) e o mCSM (0,577) obtiveram os piores resultados. A Figura 39 mostrou, através das matrizes de confusão, uma distribuição maior de acertos para a classe "Desestabilizante", isso em ambos os preditores. Para a classe "Estabilizante" os resultados comparados foram similares, porém com um número baixo de acertos. A classe "Neutra" obteve uma taxa de acerto de 76,63% pelo *I-Mutant* e 68,84% pelo EN-MUTATE: *Plurality voting*.

Os próximos experimentos abordados neste capítulo utilizaram abordagens diferentes, visto que o objetivo é comparar os diversos modelos preditivos integrantes da metodologia *ensemble* EN-MUTATE com as ferramentas de predição adotadas.

7.4 Experimento 2: conjunto de dados M3432B avaliado com os dados de treinamento

A Tabela 23 exibe os resultados obtidos com o conjunto de dados M3432B (3432 instâncias). Os melhores resultados comparados (EN-MUTATE e ferramentas de predição) baseados na Medida-F, estão destacados em negrito. A classe ($\Delta\Delta G$) foi discretizada em duas categorias: desestabilizante (2217 instâncias) e estabilizante (1215 instâncias). Complementando as análises, a Figura 40 ilustra as matrizes de confusão dos melhores preditores, EN-MUTATE: *RandomForest* seguido da ferramenta mCSM. Neste experimento, a avaliação do EN-MUTATE foi realizada através de seus modelos preditivos avaliados com os dados de treinamento pelo WEKA (*Use training set*). O mesmo conjunto de mutações foi submetido para as ferramentas de predição comparadas.

Tabela 23: Experimento 2: conjunto M3432B avaliado com os dados de treinamento.

Classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
J48	66,95%	0,714	0,670	0,569
SVM	67,36%	0,725	0,674	0,577
MultilayerPerceptron	64,59%	0,417	0,646	0,507
NaiveBayes	64,07%	0,603	0,641	0,593
Ensemble de classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
RandomForest	**98,48%**	**0,985**	**0,985**	**0,985**
Bagging (J48)	78,29%	0,820	0,783	0,755
Bagging (SVM)	69,17%	0,721	0,692	0,621
Bagging (*MultilayerPerceptron*)	64,59%	0,417	0,646	0,507
Bagging (*NaiveBayes*)	63,75%	0,599	0,638	0,592
Boosting (J48)	66,95%	0,714	0,670	0,569
Boosting (SVM)	70,54%	0,693	0,705	0,686
Boosting (*MultilayerPerceptron*)	64,59%	0,417	0,646	0,507
Boosting (*NaiveBayes*)	64,07%	0,603	0,641	0,593
Stacking [Meta classificador SVM] (J48, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	64,59%	0,417	0,646	0,507
Vote [Votação majoritária] (J48, SVM, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	65,64%	0,776	0,656	0,531
Ferramentas				
Preditor	Acurácia	Precisão	Revocação	Medida-F
I-Mutant	63,20%	0,829	0,632	0,703
CUPSAT	55,77%	0,588	0,558	0,571
SDM	53,47%	0,527	0,535	0,530
mCSM	**63,26%**	**0,838**	**0,633**	**0,707**
DUET	61,66%	0,759	0,617	0,668
iRDP	55,07%	0,584	0,551	0,565
MAESTRO	56,15%	0,616	0,561	0,584

```
EN-MUTATE: RandomForest

=== Matriz de Confusão ===

   a    b   <-- classificado como
1193   22 |    a = Estabilizante
  30 2187 |    b = Desestabilizante

(a)
```

```
Ferramenta: mCSM

=== Matriz de Confusão ===

   a    b   <-- classificado como
 143  189 |    a = Estabilizante
1072 2028 |    b = Desestabilizante

(b)
```

Figura 40: Experimento 2: Matrizes de confusão dos preditores com melhores resultados.

A predição do conjunto de dados M3432B pelos algoritmos adotados na metodologia EN-MUTATE, em geral, obteve bons resultados quando comparados as ferramentas de predição participantes do *ensemble*. Por exemplo, o preditor *RandomForest* (0,985) obteve o melhor valor de Medida-F seguido do *Bagging* (J48) (0,755). As melhores ferramentas, mCSM e *I-Mutant*, alcançaram valores de Medida-F de 0,707 e 0,703, respectivamente. É importante mencionar que o uso da técnica de *Bagging* sobre os preditores do EN-MUTATE, principalmente nos casos do J48 e SVM, resultou em um aumento considerável de Medida-F. No entanto, foi observado que o tamanho da árvore do algoritmo J48 passou de 13 para 223, sendo que árvores grandes tendem a ser muito especializadas. Conforme apresentou a Figura 40 (a), o *RandomForest* acertou 1193 instâncias da classe "Estabilizante" e 2187 instâncias da classe "Desestabilizante". Também foi constatado que o mCSM, Figura 40 (b), teve um forte viés em predizer a classe "Desestabilizante" (2028 acertos), no entanto, teve somente 143 acertos de um total de 1215 instâncias da classe "Estabilizante".

7.5 Experimento 3: conjunto de dados M3432B avaliado com um conjunto de teste

Os próximos valores apresentados pela Tabela 24 se referem aos resultados aplicados ao conjunto de dados M3432B (3432 instâncias), porém, avaliado com um conjunto de teste contendo 1029 instâncias (30% do total) subtraídas aleatoriamente e não utilizadas para treinamento dos modelos EN-MUTATE. Os dois melhores resultados comparados (EN-MUTATE e ferramentas de predição) estão destacados em negrito e são igualmente apresentados na Figura 41 através de matrizes de confusão. Nesse experimento, a classe ($\Delta\Delta G$) foi discretizada em duas categorias: desestabilizante (1552 instâncias para treinamento e 665 para teste) e estabilizante (851 instâncias para treinamento e 364 para teste).

Tabela 24: Experimento 3: conjunto M3432B avaliado com um conjunto de teste.

Classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
J48	**64,82%**	**0,612**	**0,648**	**0,595**
SVM	63,55%	0,573	0,636	0,549
MultilayerPerceptron	64,62%	0,418	0,646	0,507
NaiveBayes	62,58%	0,582	0,626	0,579
Ensemble de classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
RandomForest	62,19%	0,579	0,622	0,579
Bagging (J48)	64,23%	0,598	0,642	0,574
Bagging (SVM)	63,94%	0,580	0,639	0,547

Bagging (*MultilayerPerceptron*)	64,62%	0,418	0,646	0,507
Bagging (*NaiveBayes*)	62,01%	0,574	0,620	0,574
Boosting (J48)	65,88%	0,645	0,659	0,567
Boosting (SVM)	59,86%	0,570	0,599	0,577
Boosting (*MultilayerPerceptron*)	64,62%	0,418	0,646	0,507
Boosting (*NaiveBayes*)	62,58%	0,582	0,626	0,579
Stacking [Meta classificador SVM] (J48, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	64,62%	0,418	0,646	0,507
Vote [Votação majoritária] (J48, SVM, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	64,62%	0,598	0,646	0,535
Ferramentas				
Preditor	Acurácia	Precisão	Revocação	Medida-F
I-Mutant	**63,75%**	**0,840**	**0,638**	**0,710**
CUPSAT	58,60%	0,630	0,586	0,604
SDM	54,81%	0,541	0,548	0,544
mCSM	63,17%	0,824	0,632	0,701
DUET	61,32%	0,755	0,613	0,665
iRDP	55,10%	0,576	0,551	0,562
MAESTRO	57,24%	0,636	0,572	0,597

```
EN-MUTATE: J48

=== Matriz de Confusão ===

 a   b   <-- classificado como
70 294 |   a = Estabilizante
68 597 |   b = Desestabilizante

(a)
```

```
Ferramenta: I-Mutant

=== Matriz de Confusão ===

 a   b   <-- classificado como
46 318 |   a = Estabilizante
55 610 |   b = Desestabilizante

(b)
```

Figura 41: Experimento 3: Matrizes de confusão dos preditores com melhores resultados.

Quando submetido a um conjunto de teste (*supplied test set*), os preditores EN-MUTATE obtiveram valores de Medida-F entre 0,507 e 0,595, sendo que o melhor resultado se deu pelo algoritmo J48. Como pôde ser visto na Tabela 24, os resultados de ambos os algoritmos, nesse caso, não atingiram ganhos significativos com o uso de *ensemble* de classificadores (*Bagging*, *Boosting*, *Stacking* e *Vote*). Dentre as ferramentas de predição, o *I-Mutant* (0,710) e o mCSM (0,701) obtiveram os melhores valores de Medida-F. É importante destacar que tais ferramentas podem ter tido vantagem na comparação com os resultados do EN-MUTATE pelo fato de que não é possível determinar se os dados de teste foram utilizados no treinamento de seus modelos preditivos, dado que ambas também utilizam resultados experimentais do *ProTherm*.

A distribuição dos acertos exibidos na Figura 41 (a), relativos ao conjunto de dados M3432B (conjunto de teste), mostraram que o EN-MUTATE: J48 classificou corretamente mais instâncias da classe "Desestabilizante" (597 instâncias) do que da classe "Estabilizante" (70 instâncias). O mesmo aconteceu para o *I-Mutant*, Figura 41 (b), na qual acertou 610 instâncias da classe "Desestabilizante", porém, apenas 46 da classe "Estabilizante". Desta forma, ambos os preditores obtiveram uma baixa acurácia da classe "Estabilizante". Consequentemente, as amostras pertencentes à classe prevalecente (mutações desestabilizantes) levaram a uma maior precisão, mesmo que o desempenho tenha sido ruim para a classe minoritária (mutações estabilizadoras).

7.6 Experimento 4: conjunto de dados M3432T avaliado com os dados de treinamento

A Tabela 25 apresenta os resultados obtidos com o conjunto de dados M3432T (3432 instâncias). Os melhores preditores comparados (EN-MUTATE e ferramentas de predição) estão destacados em negrito. Em seguida, a Figura 42 exibe as matrizes de confusão desses mesmos preditores. Neste experimento, a classe ($\Delta\Delta G$) foi discretizada em três categorias: desestabilizante (1885 instâncias), neutra (940 instâncias) e estabilizante (607 instâncias). A avaliação dos modelos que compõem o EN-MUTATE foi realizada com os dados de treinamento (opção *Use training set* do WEKA). Assim como nos demais experimentos, também foram feitas submissões via interface *web* para as ferramentas de predição utilizando o mesmo conjunto de mutações avaliado no EN-MUTATE.

Tabela 25: Experimento 4: conjunto M3432T avaliado com os dados de treinamento.

Classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
J48	61,68%	0,647	0,617	0,562
SVM	57,77%	0,713	0,578	0,453
MultilayerPerceptron	55,41%	0,440	0,554	0,425
NaiveBayes	54,42%	0,503	0,544	0,450
Ensemble de classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
RandomForest	**98,39%**	**0,984**	**0,984**	**0,984**
Bagging (J48)	93,24%	0,937	0,932	0,931
Bagging (SVM)	58,71%	0,666	0,587	0,481
Bagging (*MultilayerPerceptron*)	54,98%	0,576	0,555	0,391
Bagging (*NaiveBayes*)	54,74%	0,506	0,547	0,451
Boosting (J48)	69,20%	0,712	0,692	0,667
Boosting (SVM)	58,44%	0,589	0,584	0,502
Boosting (*MultilayerPerceptron*)	55,41%	0,440	0,554	0,425
Boosting (*NaiveBayes*)	54,42%	0,503	0,544	0,450

Stacking [Meta classificador SVM] (J48, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	54,92%	0,302	0,549	0,389
Vote [Votação majoritária] (J48, SVM, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	58,50%	0,736	0,585	0,465
Ferramentas				
Preditor	Acurácia	Precisão	Revocação	Medida-F
I-Mutant	**47,00%**	**0,427**	**0,470**	**0,440**
CUPSAT	43,74%	0,431	0,437	0,433
SDM	37,50%	0,409	0,375	0,388
mCSM	47,12%	0,413	0,471	0,427
DUET	44,70%	0,405	0,447	0,417
iRDP	41,61%	0,412	0,416	0,413
MAESTRO	42,07%	0,406	0,421	0,409

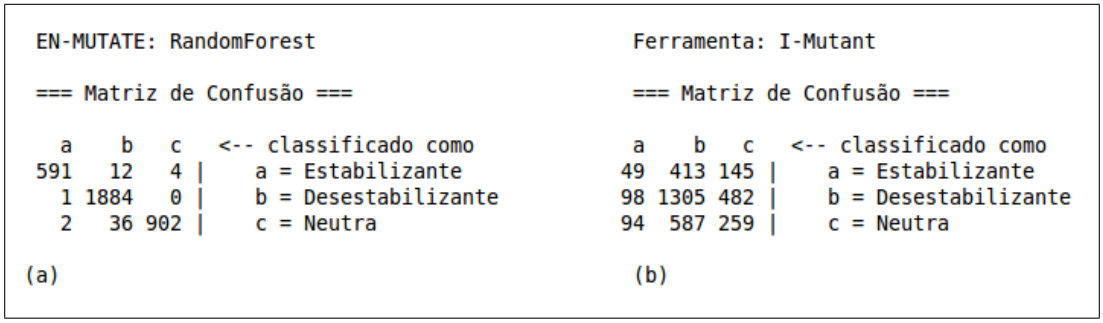


Figura 42: Experimento 4: Matrizes de confusão dos preditores com melhores resultados.

Os valores exibidos na Tabela 25 resumem os resultados do quarto experimento. Assim como no Experimento 2, os melhores valores de Medida-F foram obtidos com os classificadores EN-MUTATE: *RandomForest* (0,984) e EN-MUTATE: *Bagging* (J48) (0,931). Em contrapartida, nenhuma das sete ferramentas de predição integradas alcançaram um valor de Medida-F superior a 0,440 (*I-Mutant*). Um ponto que deve ser mencionado é que, da mesma forma para o conjunto de dados M3432B, quando adotada a técnica de *Bagging* ao J48, houve uma significativa melhora de Medida-F em relação ao seu uso padrão, entretanto, o tamanho da árvore passou, originalmente, de 159 para 900 (superajustada aos dados de treinamento).

Ainda quanto aos resultados do classificador EN-MUTATE: *RandomForest* (Figura 42), visualizados através da diagonal principal da sua matriz de confusão, pôde ser observado uma boa quantidade de acertos para ambas as classes: "Desestabilizante" (1884 instâncias), "Estabilizante" (591 instâncias) e "Neutra" (906 instâncias). Um outro ponto apresentado pela matriz de confusão, na qual colaborou para o baixo valor de Medida-F do *I-Mutant* (melhor resultado), foi que a ferramenta classificou 587 mutações da classe

"Desestabilizante" como sendo "Neutra". Devido a proximidade dos valores que constituem essas duas classes, o erro pode ser considerado menos prejudicial, entretanto, o baixo número de acertos da classe "Estabilizante" (49 instâncias) acabou sendo o fator mais agravante nesse caso.

7.7 Experimento 5: conjunto de dados M3432T avaliado com um conjunto de teste

Finalizando os resultados obtidos, a Tabela 26 apresenta os detalhes da acurácia de predição dos classificadores que integram a metodologia EN-MUTATE, assim como das ferramentas de predição comparadas. Os preditores com os melhores valores de Medida-F (EN-MUTATE e ferramentas de predição) estão destacados em negrito bem como exibidos através de suas matrizes de confusão (Figura 43). Neste experimento, foi utilizado o conjunto de dados M3432T (3432 instâncias), na qual foi subtraído um conjunto de teste contendo 1029 instâncias (30% do total) escolhidas aleatoriamente e não utilizadas durante o treinamento dos modelos EN-MUTATE. Esse mesmo conjunto de mutações foi submetido para as ferramentas de predição. A classe ($\Delta\Delta G$) foi discretizada em três categorias: desestabilizante (1318 instâncias para treinamento e 567 para teste), neutra (658 instâncias para treinamento e 282 para teste) e estabilizante (427 instâncias para treinamento e 180 para teste).

Tabela 26: Experimento 5: conjunto M3432T avaliado com um conjunto de teste.

Classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
J48	53,25%	0,447	0,533	0,467
SVM	54,90%	0,451	0,549	0,403
MultilayerPerceptron	54,51%	0,403	0,545	0,414
NaiveBayes	53,45%	0,444	0,534	0,457
Ensemble de classificadores				
Preditor	Acurácia	Precisão	Revocação	Medida-F
RandomForest	50,53%	0,459	0,505	0,469
***Bagging* (J48)**	**51,79%**	**0,478**	**0,518**	**0,482**
Bagging (SVM)	54,71%	0,480	0,547	0,415
Bagging (*MultilayerPerceptron*)	55,10%	0,433	0,551	0,405
Bagging (*NaiveBayes*)	53,15%	0,444	0,532	0,458
Boosting (J48)	53,25%	0,447	0,533	0,467
Boosting (SVM)	53,54%	0,441	0,535	0,439
Boosting (*MultilayerPerceptron*)	54,51%	0,403	0,545	0,414
Boosting (*NaiveBayes*)	53,45%	0,444	0,534	0,457

Stacking [Meta classificador SVM] (J48, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	55,10%	0,304	0,551	0,392
Vote [Votação majoritária] (J48, SVM, *RandomForest*, *MultilayerPerceptron*, *NaiveBayes*)	55,39%	0,494	0,554	0,421
Ferramentas				
Preditor	Acurácia	Precisão	Revocação	Medida-F
I-Mutant	46,36%	0,418	0,464	0,434
CUPSAT	**44,02%**	**0,437**	**0,440**	**0,438**
SDM	37,41%	0,409	0,374	0,387
mCSM	47,72%	0,412	0,477	0,430
DUET	44,51%	0,400	0,445	0,414
iRDP	42,18%	0,413	0,422	0,417
MAESTRO	41,50%	0,395	0,415	0,402

```
EN-MUTATE: Bagging (J48)

=== Matriz de Confusão ===

  a   b   c   <-- classificado como
 24 112  44 |   a = Estabilizante
 29 433 150 |   b = Desestabilizante
 16 190  76 |   c = Neutra

(a)
```

```
Ferramenta: CUPSAT

=== Matriz de Confusão ===

  a   b   c   <-- classificado como
 33 103  44 |   a = Estabilizante
 98 333 136 |   b = Desestabilizante
 49 146  87 |   c = Neutra

(b)
```

Figura 43: Experimento 5: Matrizes de confusão dos preditores com melhores resultados.

No quinto experimento realizado (Tabela 26), quando aplicado um conjunto de teste aos modelos e igualmente submetido as ferramentas de predição, a metodologia EN-MUTATE manteve os resultados mais acurados. Desta vez, o algoritmo J48 (utilizando a técnica *Bagging*) obteve o maior valor de Medida-F (0,482). O pior resultado se deu pelo algoritmo *Stacking* (Medida-F = 0,392). Dentre as ferramentas de predição, o CUPSAT obteve o melhor resultado (Medida-F = 0,438), sendo que o pior resultado foi obtido pelo SDM (Medida-F = 0,387).

A matriz de confusão do classificador EN-MUTATE *Bagging*: J48, representada na Figura 43 (a), indica que o maior número de acertos se concentrou na classe majoritária "Desestabilizante" (433 instâncias). Em contrapartida, apenas 24 instâncias da classe "Estabilizante" foram classificadas corretamente. Igualmente, pôde ser visto que uma grande quantidade de instâncias da classe "Neutra" foram classificadas como "Desestabilizante". No que se refere a matriz de confusão do CUPSAT, Figura 43 (b), mesmo com a Acurácia e Medida-F menores que a do EN-MUTATE *Bagging*: J48, a ferramenta apresentou uma melhor distribuição dos seus resultados, acertando mais instâncias das classes "De-

sestabilizante" e "Neutra". Os acertos da classe "Estabilizante" pelo CUPSAT também foram baixos (33 instâncias) e 146 instâncias da classe "Neutra" foram classificadas como "Desestabilizante".

7.8 Discussões finais

Este capítulo apresentou os resultados obtidos através dos testes de hipóteses dos modelos preditivos baseados na proposta EN-MUTATE em comparação com as ferramentas individuais que fizeram parte do *ensemble* adotado. Os conjuntos de dados foram compostos por valores experimentais termodinâmicos de mutações pontuais disponibilizados pela base de dados *ProTherm*. A opção de validação cruzada tipicamente aplicada a modelos preditivos não foi considerada nos experimentos pelo fato de que o principal objetivo é comparar o uso de *ensemble learning* através da metodologia proposta com a predição individual de cada ferramenta descrita no Capítulo 4. Consequentemente, não é possível reproduzir fielmente seus conjuntos de dados utilizados tanto para treinamento quanto para teste, pois o conteúdo é limitado ao apresentado nas publicações. Isso de fato pode gerar uma certa vantagem às ferramentas comparadas, uma vez que grande parte delas utilizou dados do *ProTherm* na definição de seus modelos. Para a validação dos modelos EN-MUTATE, os experimentos utilizaram as opções "*Use training set*" e "*Supplied test set*" do WEKA. É importante mencionar que foram usados os mesmos conjuntos de mutações para as avaliações de predição das ferramentas comparadas.

Foi constatado um baixo índice de instâncias classificadas corretamente pelo EN-MUTATE para as mutações pontuais estabilizantes, o que se repetiu também para as ferramentas de predição. Segundo Worth, Preissner, Blundell (2011), a maioria das mutações encontradas na natureza são desestabilizantes e isso se reflete nos conjuntos de dados termodinâmicos mutantes usados para desenvolver e testar os métodos preditivos. Um fator que pode ter contribuído para a diferença nos resultados das variações de $\Delta\Delta G$ é um desbalanceamento entre as mutações pontuais desestabilizantes, neutras e estabilizantes durante o treinamento e definição dos modelos das ferramentas adotadas nas tarefas de *Ensemble Learning*.

A existência de classes com uma quantidade significativamente maior de instâncias que as demais pode levar à indução de classificadores tendenciosos para classes majoritárias (FACELI et al., 2011). Em problemas de classificação, um conjunto de dados é definido como desequilibrado quando o número de instâncias que representa uma classe é menor do que os de outras classes (GALAR et al., 2012). Além disso, do ponto de vista da tarefa de aprendizado, a classe com o menor número de casos é geralmente a classe de interesse (YANG et al., 2014). Isso efetivamente é um grande desafio na implementação de novas abordagens para o problema de predição do impacto de mutações pontuais em proteínas e também foi discutido no trabalho de Malinka (2015), no qual o autor evidencia

que os valores preditos por grande parte das ferramentas de predição são tendenciosos em direção a valores de $\Delta\Delta G$ negativos, embora a ferramenta SDM tenha sido mais precisa nas mutações estabilizantes avaliadas nos experimentos.

Da mesma forma, as observações de Chen, Lin e Chu (2013) verificaram que as ferramentas por eles avaliadas também realizaram mais predições da classe "Desestabilizante" do que "Estabilizante", levando a uma alta especificidade (taxa de acerto da classe negativa), entretanto com uma baixa revocação (taxa de acerto da classe positiva). A partir disso, é provável que a maior parte dos aminoácidos usados nas mutações que formaram os conjuntos de dados causaram desestabilização da proteína, além disso, do ponto de vista biológico, mutações são mudanças no curso normal da natureza, o que aumenta a complexidade da criação de um modelo preditivo.

Finalizando as discussões finais, a Tabela 27 reúne os melhores resultados obtidos em ambos os cinco experimentos realizados nesta dissertação. Os preditores com o maiores valores de Medida-F estão destacados em negrito.

Tabela 27: Resumo dos melhores resultados obtidos nos experimentos.

Experimento 1				
Preditor	Acurácia	Precisão	Revocação	Medida-F
Plurality voting	73,13%	0,740	0,731	0,721
I-Mutant	**76,06%**	**0,763**	**0,761**	**0,753**
Experimento 2				
Preditor	Acurácia	Precisão	Revocação	Medida-F
RandomForest	**98,48%**	**0,985**	**0,985**	**0,985**
mCSM	63,26%	0,838	0,633	0,707
Experimento 3				
Preditor	Acurácia	Precisão	Revocação	Medida-F
J48	64,82%	0,612	0,648	0,595
I-Mutant	**63,75%**	**0,840**	**0,638**	**0,710**
Experimento 4				
Preditor	Acurácia	Precisão	Revocação	Medida-F
RandomForest	**98,39%**	**0,984**	**0,984**	**0,984**
I-Mutant	47,00%	0,427	0,470	0,440
Experimento 5				
Preditor	Acurácia	Precisão	Revocação	Medida-F
***Bagging* (J48)**	**51,79%**	**0,478**	**0,518**	**0,482**
CUPSAT	44,02%	0,437	0,440	0,438

8 CONCLUSÃO

A expectativa de desenvolver processos que permitam localizar e mapear as mutações ocorridas em proteínas nos possibilita um diagnóstico mais acurado e, do mesmo modo, proporciona uma maior compreensão dos eventos moleculares (VOET; VOET; PRATT, 2014). Com isso, uma aplicação que prediz a estabilidade térmica de mutantes pode ser útil para a condução do processo de tomada de decisão na modelagem de uma proteína através de mutagênese (JIA; YARLAGADDA; REED, 2015).

Na Bioinformática em si, diferentes desafios de implementação são enfrentados tendo em vista que o ambiente computacional não consegue simular com perfeição o que acontece na natureza, mesmo assim, novas propostas tentam suprir essa necessidade. Por esse motivo, a combinação de métodos computacionais e estatísticos com dados experimentais, incluindo novas tecnologias como o sequenciamento profundo e medições de estabilidade em alta precisão, apresentam muitas abordagens de apoio para a engenharia de estabilidade de proteínas, mesmo com os inerentes problemas computacionais existentes (MAGLIERY, 2015). Nesse cenário, os métodos de AM podem ter a sua aplicação bem sucedida, uma vez que servem de instrumento aos especialistas na busca intrínseca por informações sobre proteínas. Do mesmo modo, são capazes de inferir modelos usados para resolver tarefas de análise de dados (SOMMER; GERLICH, 2013).

Uma boa avaliação do desempenho de um método de AM precisa de um conjunto de dados abrangente e representativo o bastante para o objetivo específico, pois em função da diversidade de tipos de dados e tarefas de análise em bioinformática, muitas vezes é difícil estimar o desempenho de métodos de aprendizado publicados com base nos dados específicos de prova de conceito utilizados no respectivo estudo (SOMMER; GERLICH, 2013). Atualmente, uma das únicas fontes públicas de resultados experimentais termodinâmicos referentes a mudanças de energia livre ($\Delta\Delta G$) é o *ProTherm* (BAVA et al., 2004), justificado pelo grande número de citações na literatura. Assim, o conteúdo do *ProTherm* pode não ser considerado como representativo para algumas abordagens, uma vez que em comparação com *Protein Data Bank* (banco de dados com informações estruturais), o *ProTherm* tem o seu crescimento demasiadamente lento (LAIMER et al., 2015). Consequentemente, as abordagens encontradas na literatura que não utilizam essa base pública

como principal fonte de dados, acabam apresentando estudos de caso mais específicos e focados em um pequeno grupo de proteínas (PANIGRAHI et al., 2015; FARISELLI et al., 2015; WITVLIET et al., 2016).

Quando se trata de combinar múltiplos modelos em conjunto para produzir *Ensemble Learning*, uma analogia que pode ser feita é com as comissões compostas de seres humanos (SUROWIECKI, 2005). Um modelo *ensemble* representa uma hipótese, contudo, essa hipótese não está necessariamente contida no espaço de hipóteses dos modelos base a partir dos quais ela é construída (FACELI et al., 2011). Em sistemas biológicos, esses modelos são frequentemente construídos a partir de conjuntos que são derivados de dados experimentais (MARDER; TAYLOR, 2011). Diante disso, é necessário discutir como medir e otimizar o desempenho de um classificador, o que pode ser útil para permitir estratégias de otimização específicas (SOMMER; GERLICH, 2013).

De forma a mensurar a viabilidade de aplicação do *ensemble* apresentado, esta dissertação avaliou seus resultados com base em valores biológicos experimentais. As primeiras versões da metodologia proposta, EN-MUTATE, realizaram o *ensemble* por meio de uma votação por pluralidade entre as ferramentas integradas. O desenvolvimento dessa primeira etapa em conjunto com a investigação das ferramentas correlatas foram importantes para se entender os mecanismos de funcionamento das abordagens de predição do impacto de mutações pontuais em proteínas. À vista disso, com a necessidade de se expandir as análises com o intuito de permitir uma metodologia baseada em modelos treinados através de diferentes classificadores, a abordagem proposta foi reestruturada e passou a abordar múltiplas opções de predição *ensemble*, o que acabou sendo agregado a ferramenta desenvolvida EN-MUTATE*web*.

Como apresentaram as Tabelas 23, 24, 25 e 26, tanto para o conjunto de dados M3432B (classes binárias) quanto para o M3432T (classes ternárias), a metodologia EN-MUTATE obteve, em grande parte, modelos mais acurados. Desse modo, as principais contribuições obtidas com o desenvolvimento desta dissertação atendem ao seu principal objetivo: definir uma metodologia cuja finalidade é adotar o conceito de *Ensemble Learning* para combinar em uma única abordagem os resultados de diferentes ferramentas de predição do impacto de mutações pontuais em proteínas, buscando, assim, a adoção de abordagens para produzir um resultado final em conjunto potencialmente melhor do que os individuais.

Como novos desafios à esta pesquisa, novas implementações mais sofisticadas podem ser exploradas para a descoberta de conhecimento sobre o impacto de mutações pontuais em proteínas, especialmente, no que se refere ao uso dos classificadores mais acurados desta dissertação, no entanto, em abordagens que fazem o uso de diferentes tipos de informação biológica. Além disso, como trabalhos futuros podem ser elencados:

- **Aplicação de novos mecanismos de combinação**: propor uma abordagem de predição que objetiva um *ensemble* aplicado a novas combinações de resultados;

- **Uso de técnicas de AM para lidar com desbalanceamento de classes**: explorar técnicas específicas para o desbalanceamento, como o *up-sampling* ou *down-sampling* (WANG; MINKU; YAO, 2013), SMOTE (CHAWLA et al., 2002) e classificação com custo (PELAYO; DICK, 2012);

- **Melhorias na acurácia de predição para as mutações estabilizantes**: nesta dissertação foi realizada uma abordagem *in-silico* sem o conhecimento prévio em relação às proteínas em estudo. Uma outra alternativa seria a definição de uma metodologia para filtrar a seleção das moléculas de acordo com parâmetros biológicos específicos, diminuindo o espaço de busca através da adição de conhecimento;

- **Novas opções para o EN-MUTATE*web***: a integração com ferramentas de visualização de estrutura de proteína que evidenciem o impacto de uma mutação pontual, possibilitando, deste modo, que a análise do mutante seja contemplada tridimensionalmente. A inclusão de novas ferramentas de predição ao *ensemble* proposto também é uma das principais motivações para a ferramenta desenvolvida;

- **Realização de estudos de caso baseados em Beta-glicosidases específicas**: sugerir mutações para a manipulação genética de algas produtoras de Beta-glicosidases com características catalíticas eficientes bem como de tolerância à inibição por glicose e celobiose, permitindo o seu uso industrial na produção de etanol de segunda geração, contribuindo, desta forma, para o projeto de pesquisa na qual este trabalho está vinculado.

REFERÊNCIAS

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine learning**, v.6, n.1, p.37–66, 1991.

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Biologia molecular da célula**. Artmed Editora, 2009.

ALENCAR, S. A. **Utilização de ferramentas computacionais para o estudo do impacto funcional e estrutural de nsSNPs em genes codificadores de proteínas**. 2010. 128 f. Tese (Doutorado em Bioinformática). Universidade Federal de Minas Gerais, Belo Horizonte.

ANDREINI, C.; CAVALLARO, G.; LORENZINI, S. FindGeo: a tool for determining metal coordination geometry. **Bioinformatics**, v.28, n.12, p.1658–1660, 2012.

APWEILER, R.; BAIROCH, A.; WU, C. H.; BARKER, W. C.; BOECKMANN, B.; FERRO, S.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M. et al. UniProt: the universal protein knowledgebase. **Nucleic acids research**, v.32, n.suppl 1, p.D115–D119, 2004.

BARRETT, G. **Chemistry and biochemistry of the amino acids**. Springer Science & Business Media, 2012.

BAVA, K. A.; GROMIHA, M. M.; UEDAIRA, H.; KITAJIMA, K.; SARAI, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. **Nucleic acids research**, v.32, n.suppl 1, p.D120–D121, 2004.

BERLINER, N.; TEYRA, J.; ÇOLAK, R.; LOPEZ, S. G.; KIM, P. M. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. **PloS one**, v.9, n.9, p.e107353, 2014.

BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P. E. The Protein Data Bank. **Nucleic acids research**, v.28, n.1, p.235–242, 2000.

BEROUD, C.; SOUSSI, T. The UMD-p53 database: new mutations and analysis tools. **Human mutation**, v.21, n.3, p.176–181, 2003.

BETTELHEIM, F.; BROWN, W.; CAMPBELL, M.; FARRELL, S. Introdução à Química Geral, Orgânica e Bioquímica. trad. **Cengage Learning São Paulo**, 2012.

BISHOP, C. M. **Pattern recognition and Machine Learning**. New York, NY: Springer, 2006.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: COMPUTATIONAL LEARNING THEORY, 1992. **Proceedings...** 1992. p.144–152.

BREIMAN, L. Bagging predictors. **Machine learning**, v.24, n.2, p.123–140, 1996.

BREIMAN, L. Random forests. **Machine learning**, v.45, n.1, p.5–32, 2001.

CAPRIOTTI, E.; FARISELLI, P.; CASADIO, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. **Nucleic acids research**, v.33, n.suppl 2, p.W306–W310, 2005.

CAPRIOTTI, E.; FARISELLI, P.; ROSSI, I.; CASADIO, R. A three-state prediction of single point mutations on protein stability changes. **BMC bioinformatics**, v.9, n.2, p.1, 2008.

CHANG, C.-C.; LIN, C.-J. LIBSVM: a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v.2, n.3, p.27, 2011.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v.16, p.321–357, 2002.

CHEN, C.-W.; LIN, J.; CHU, Y.-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. **BMC bioinformatics**, v.14, n.2, p.1, 2013.

CHEN, H.; POON, J.; POON, S. K.; CUI, L.; FAN, K.; SZE, D. M. Ensemble learning for prediction of the bioactivity capacity of herbal medicines from chromatographic fingerprints. **BMC bioinformatics**, v.16, n.Suppl 12, p.S4, 2015.

CHENG, J.; RANDALL, A.; BALDI, P. Prediction of protein stability changes for single-site mutations using support vector machines. **Proteins: Structure, Function, and Bioinformatics**, v.62, n.4, p.1125–1132, 2006.

DANUSER, G. Computer vision in cell biology. **Cell**, v.147, n.5, p.973–978, 2011.

DASARATHY, B. V.; SHEELA, B. V. A composite classifier system design: concepts and methodology. **Proceedings of the IEEE**, v.67, n.5, p.708–713, 1979.

DEHOUCK, Y.; GROSFILS, A.; FOLCH, B.; GILIS, D.; BOGAERTS, P.; ROOMAN, M. PoPMuSiC-2.0: Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks. **Bioinformatics**, v.25, n.19, p.2537–2543, 2009.

DIAS, S. R. **Residue Interaction Database – Proposição de mutações sítio dirigidas com base em interações observadas em proteínas de estrutura tridimensional conhecida**. 2012. Tese (Doutorado em Bioinformática). 119 f. Universidade Federal de Minas Gerais, Belo Horizonte.

DIETTERICH, T. G. Ensemble methods in machine learning. In: **Multiple classifier systems**. Springer, 2000. p.1–15.

DURHAM, E. H. A. B. **Bioinformática Estrutural de Proteínas Modificadas por Eventos de Splicing Alternativo**. 2007. Tese (Doutorado em Ciências). 131 f. Universidade de São Paulo, São Paulo.

DZEROSKI, S.; ZENKO, B. Is combining classifiers with stacking better than selecting the best one? **Machine learning**, v.54, n.3, p.255–273, 2004.

EBRAHIMPOUR, R.; SADEGHNEJAD, N.; AMIRI, A.; MOSHTAGH, A. Low resolution face recognition using combination of diverse classifiers. In: SOFT COMPUTING AND PATTERN RECOGNITION (SOCPAR), 2010 INTERNATIONAL CONFERENCE OF, 2010. **Anais...** 2010. p.265–268.

EDDY, S. R. Accelerated profile HMM searches. **PLoS Comput Biol**, v.7, n.10, p.e1002195, 2011.

EICKHOLT, J.; CHENG, J. DNdisorder: predicting protein disorder using boosting and deep networks. **BMC bioinformatics**, v.14, n.1, p.1, 2013.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. **Inteligência Artificial – Uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011.

FARISELLI, P.; MARTELLI, P. L.; SAVOJARDO, C.; CASADIO, R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. **Bioinformatics**, v.31, n.17, p.2816–2821, 2015.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. AAAI press Menlo Park, 1996. v.21.

FISER, A.; SALI, A. Modeller: generation and refinement of homology-based protein structure models. **Methods in enzymology**, v.374, p.461–491, 2003.

FOLKMAN, L.; STANTIC, B.; SATTAR, A. Feature-based multiple models improve classification of mutation-induced stability changes. **BMC genomics**, v.15, n.4, p.S6, 2014.

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: ICML, 1996. **Anais...** 1996. v.96, p.148–156.

GALAR, M.; FERNANDEZ, A.; BARRENECHEA, E.; BUSTINCE, H.; HERRERA, F. A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v.42, n.4, p.463–484, 2012.

GAMA, J.; BRAZDIL, P. Cascade generalization. **Machine Learning**, v.41, n.3, p.315–343, 2000.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. **Atmospheric environment**, v.32, n.14, p.2627–2636, 1998.

GETOV, I.; PETUKH, M.; ALEXOV, E. SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. **International journal of molecular sciences**, v.17, n.4, p.512, 2016.

GOUGH, B. **GNU scientific library reference manual**. Network Theory Ltd., 2009.

HAN, J.; PEI, J.; KAMBER, M. **Data mining – concepts and techniques**. San Francisco: Morgan and Kaufmann, 2006.

HANSEN, L. K.; SALAMON, P. Neural network ensembles. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, n.10, p.993–1001, 1990.

HARVEY, R.; FERRIER, D. Lippincott's illustrated reviews – Biochemistry. **Edition by Kluwer, New York**, 2011.

HASTIE, T. J.; TIBSHIRANI, R. J.; FRIEDMAN, J. H. **The elements of statistical learning – data mining, inference, and prediction**. Springer, 2011.

HO, T. K. Multiple classifier combination: Lessons and next steps. In: **Hybrid methods in pattern recognition**. World Scientific, 2002. p.171–198.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies. **Artificial intelligence review**, v.22, n.2, p.85–126, 2004.

HOGEWEG, P. The roots of bioinformatics in theoretical biology. **PLoS Comput Biol**, v.7, n.3, p.1002021, 2011.

HORTON, H. et al. **Principios de bioquímica**. Pearson México, 2008.

HU, Q.; MERCHANTE, C.; STEPANOVA, A. N.; ALONSO, J. M.; HEBER, S. A Stacking-Based Approach to Identify Translated Upstream Open Reading Frames in Arabidopsis Thaliana. In: INTERNATIONAL SYMPOSIUM ON BIOINFORMATICS RESEARCH AND APPLICATIONS, 2015. **Anais...** 2015. p.138–149.

HUANG, L.-T.; GROMIHA, M. M.; HO, S.-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. **Bioinformatics**, v.23, n.10, p.1292–1293, 2007.

HUANG, L.-T.; WU, C.-C.; LAI, L.-F.; GROMIHA, M. M.; WANG, C.-S.; CHEN, Y.-R. Data mining application in biomedical informatics for probing into protein stability upon double mutation. **Appl. Math**, v.8, n.1L, p.125–132, 2014.

HUBBARD, S. J.; THORNTON, J. M. Naccess. **Computer Program, Department of Biochemistry and Molecular Biology, University College London**, v.2, n.1, 1993.

JIA, L.; YARLAGADDA, R.; REED, C. C. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. **PloS one**, v.10, n.9, p.e0138022, 2015.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. In: ELEVENTH CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 1995. **Proceedings...** 1995. p.338–345.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, v.22, n.12, p.2577–2637, 1983.

KARLOS, S.; FAZAKIS, N.; KOTSIANTIS, S.; SGARBAS, K. A semisupervised cascade classification algorithm. **Applied Computational Intelligence and Soft Computing**, v.2016, p.4, 2016.

KAYNAK, C.; ALPAYDIN, E. Multistage cascading of multiple classifiers: One man's noise is another man's data. In: ICML, 2000. **Anais...** 2000. p.455–462.

KHAN, S.; VIHINEN, M. Performance of protein stability predictors. **Human mutation**, v.31, n.6, p.675–684, 2010.

KITTLER, J.; HATEF, M.; DUIN, R.; MATAS, J. On combining classifiers. **IEEE transactions on pattern analysis and mach**, v.20, n.3, p.226–239, 1998.

KOHAVI, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: KDD, 1996. **Anais...** 1996. v.96, p.202–207.

KUNCHEVA, L. I. **Combining pattern classifiers – methods and algorithms**. John Wiley & Sons, 2004.

LAIMER, J.; HIEBL-FLACH, J.; LENGAUER, D.; LACKNER, P. MAESTROweb – a web server for structure based protein stability prediction. **Bioinformatics**, p.btv769, 2016.

LAIMER, J.; HOFER, H.; FRITZ, M.; WEGENKITTL, S.; LACKNER, P. MAESTRO – multi agent stability prediction upon point mutations. **BMC bioinformatics**, v.16, n.1, p.116, 2015.

LAZAR, P.; KIM, S.; LEE, Y.; SON, M.; KIM, H.-H.; KIM, Y. S.; LEE, K. W. Molecular modeling study on the effect of residues distant from the nucleotide-binding portion on RNA binding in Staphylococcus aureus Hfq. **Journal of Molecular Graphics and Modelling**, v.28, n.3, p.253–260, 2009.

LESK, A. M.; ANDRADE, A. E. **Introdução à bioinformática**. Artmed, 2008.

LIN, X.; YACOUB, S.; BURNS, J.; SIMSKE, S. Performance analysis of pattern classifier combination by plurality voting. **Pattern Recognition Letters**, v.24, n.12, p.1959–1969, 2003.

LODISH, H.; BERK, A.; KAISER, C. A.; KRIEGER, M.; BRETSCHER, A.; PLOEGH, H.; AMON, A. **Biologia celular e molecular**. Artmed Editora, 2014.

MAGLIERY, T. J. Protein stability – computation, sequence statistics, and new experimental methods. **Current opinion in structural biology**, v.33, p.161–168, 2015.

MALINKA, F. Prediction of protein stability changes upon one-point mutations using machine learning. In: CONFERENCE ON RESEARCH IN ADAPTIVE AND CONVERGENT SYSTEMS, 2015., 2015. **Proceedings...** 2015. p.102–107.

MARDER, E.; TAYLOR, A. L. Multiple models to capture the variability in biological neurons and networks. **Nature neuroscience**, v.14, n.2, p.133–138, 2011.

MASSO, M.; VAISMAN, I. I. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. **Protein Engineering Design and Selection**, p.gzq042, 2010.

MELO, F.; FEYTMANS, E. Novel knowledge-based mean force potential at atomic level. **Journal of molecular biology**, v.267, n.1, p.207–222, 1997.

MENDOZA, M. R. **Exploring ensemble learning techniques to optimize the reverse engineering of gene regulatory networks**. 2014. Tese (Doutorado em Ciência da Computação). 219 f. Universidade Federal do Rio Grande do Sul, Porto Alegre.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning – An artificial intelligence approach**. Springer Science & Business Media, 2013.

MIZUGUCHI, K.; DEANE, C. M.; BLUNDELL, T. L.; OVERINGTON, J. P. HOMSTRAD: a database of protein structure alignments for homologous families. **Protein science**, v.7, n.11, p.2469–2471, 1998.

MOAL, I. H.; FERNÁNDEZ-RECIO, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. **Bioinformatics**, v.28, n.20, p.2600–2607, 2012.

NAGI, S.; BHATTACHARYYA, D. Classification of microarray cancer data using ensemble approach. **Network Modeling Analysis in Health Informatics and Bioinformatics**, v.2, n.3, p.159–173, 2013.

NELSON, D. L.; LEHNINGER, A. L.; COX, M. M. **Lehninger principles of biochemistry**. Macmillan, 2008.

OLIVEIRA, L. S.; BRITTO, A.; SABOURIN, R. Improving cascading classifiers with particle swarm optimization. In: DOCUMENT ANALYSIS AND RECOGNITION, 2005. PROCEEDINGS. EIGHTH INTERNATIONAL CONFERENCE ON, 2005. **Anais...** 2005. p.570–574.

PANIGRAHI, P.; SULE, M.; GHANATE, A.; RAMASAMY, S.; SURESH, C. Engineering Proteins for Thermostability with iRDP Web Server. **PloS one**, v.10, n.10, p.e0139486, 2015.

PARK, S.-Y.; YOO, M.-J.; SHIN, J.-M.; CHO, K.-H. SABA (secondary structure assignment program based on only alpha carbons): a novel pseudo center geometrical criterion for accurate assignment of protein secondary structures. **BMB reports**, v.44, n.2, p.118–122, 2011.

PARTHIBAN, V.; GROMIHA, M. M.; SCHOMBURG, D. CUPSAT: prediction of protein stability upon point mutations. **Nucleic acids research**, v.34, n.suppl 2, p.W239–W242, 2006.

PELAYO, L.; DICK, S. Evaluating stratification alternatives to improve software defect prediction. **IEEE Transactions on Reliability**, v.61, n.2, p.516–525, 2012.

PENNISI, E. The human genome. **Science**, v.291, n.5507, p.1177, 2001.

PETUKH, M.; KUCUKKAL, T. G.; ALEXOV, E. On Human Disease-Causing Amino Acid Variants: Statistical Study of Sequence and Structural Patterns. **Human mutation**, v.36, n.5, p.524–534, 2015.

PIRES, D. E.; ASCHER, D. B.; BLUNDELL, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics**, v.30, n.3, p.335–342, 2014.

PIRES, D. E.; ASCHER, D. B.; BLUNDELL, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. **Nucleic acids research**, p.gku411, 2014.

PIRES, D. E.; MELO-MINARDI, R. C. de; SANTOS, M. A. dos; SILVEIRA, C. H. da; SANTORO, M. M.; MEIRA, W. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. **BMC genomics**, v.12, n.4, p.1, 2011.

PIRES, D. E.; MELO-MINARDI, R. C. de; SILVEIRA, C. H. da; CAMPOS, F. F.; MEIRA, W. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics**, v.29, n.7, p.855–861, 2013.

POLIKAR, R. Ensemble based systems in decision making. **Circuits and systems magazine, IEEE**, v.6, n.3, p.21–45, 2006.

POLIKAR, R. Ensemble learning. In: **Ensemble machine learning**. Springer, 2012. p.1–34.

QUINLAN, J. R. **C4.5**: Programs for Machine Learning. Morgan Kaufmann, 1993. v.1.

REID, M. E.; LOMAS-FRANCIS, C.; OLSSON, M. L. **The blood group antigen factsbook**. Academic Press, 2012.

RIDDER, D. de; RIDDER, J. de; REINDERS, M. J. Pattern recognition in bioinformatics. **Briefings in bioinformatics**, v.14, n.5, p.633–647, 2013.

ROKACH, L.; MAIMON, O. **Data mining with decision trees – theory and applications**. World scientific, 2014.

ROZZA, A.; LOMBARDI, G.; RE, M.; CASIRAGHI, E.; VALENTINI, G.; CAMPADELLI, P. A novel ensemble technique for protein subcellular location prediction. In: **Ensembles in Machine Learning Applications**. Springer, 2011. p.151–167.

SARABOJI, K.; GROMIHA, M. M.; PONNUSWAMY, M. Average assignment method for predicting the stability of protein mutants. **Biopolymers**, v.82, n.1, p.80–92, 2006.

SCHOLKOPF, B.; SUNG, K.-K.; BURGES, C. J.; GIROSI, F.; NIYOGI, P.; POGGIO, T.; VAPNIK, V. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. **IEEE transactions on Signal Processing**, v.45, n.11, p.2758–2765, 1997.

SCHYMKOWITZ, J.; BORG, J.; STRICHER, F.; NYS, R.; ROUSSEAU, F.; SERRANO, L. The FoldX web server: an online force field. **Nucleic acids research**, v.33, n.suppl 2, p.W382–W388, 2005.

SHAFIEI, E.; JAZAYERI-RAD, H. Improving the identification performance of an industrial process using multiple neural networks. **American Journal of Intelligent Systems**, v.2, n.4, p.40–44, 2012.

SHI, X.; BARNES, R. O.; CHEN, L.; SHAJAHAN-HAQ, A. N.; HILAKIVI-CLARKE, L.; CLARKE, R.; WANG, Y.; XUAN, J. BMRF-Net: a software tool for identification of protein interaction subnetworks by a bagging Markov random field-based method. **Bioinformatics**, p.btv137, 2015.

SOARES, S. G. **Ensemble Learning Methodologies for Soft Sensor Development in Industrial Processes**. 2015. Tese (Doutorado em Engenharia Eletrotecnica e de Computadores). 240 f. Faculdade de Ciências e Tecnologia, Coimbra.

SOMMER, C.; GERLICH, D. W. Machine learning in cell biology – teaching computers to recognize phenotypes. **J Cell Sci**, v.126, n.24, p.5529–5539, 2013.

SUROWIECKI, J. **The wisdom of crowds**. Anchor, 2005.

TAGHI, M.; NAPOLITANO DITTMAN, D. J.; KHOSHGOFTAAR, A.; FAZELPOUR, A. Select-bagging: Effectively combining gene selection and bagging for balanced bioinformatics data. In: BIOINFORMATICS AND BIOENGINEERING (BIBE), 2014 IEEE INTERNATIONAL CONFERENCE ON, 2014. **Anais...** 2014. p.413–419.

TAN, P.-N. et al. **Introduction to data mining**. Pearson Education India, 2006.

TEILUM, K.; OLSEN, J. G.; KRAGELUND, B. B. Protein stability, flexibility and function. **Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics**, v.1814, n.8, p.969–976, 2011.

TOPHAM, C. M.; MCLEOD, A.; EISENMENGER, F.; OVERINGTON, J. P.; JOHNSON, M. S.; BLUNDELL, T. L. Fragment ranking in modelling of protein structure: Conformationally constrained environmental amino acid substitution tables. **Journal of molecular biology**, v.229, n.1, p.194–220, 1993.

TOPHAM, C. M.; SRINIVASAN, N.; BLUNDELL, T. L. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. **Protein Engineering**, v.10, n.1, p.7–21, 1997.

VERLI, H. et al. **Bioinformática – Da Biologia à Flexibilidade Molecular**. Porto Alegre, 2014.

VOET, D.; VOET, J. G.; PRATT, C. W. **Fundamentos de Bioquímica – A Vida em Nível Molecular**. Artmed Editora, 2014.

VOSS, N.; GERSTEIN, M. Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly. **Journal of molecular biology**, v.346, n.2, p.477–492, 2005.

WANG, G.; DUNBRACK, R. L. PISCES: a protein sequence culling server. **Bioinformatics**, v.19, n.12, p.1589–1591, 2003.

WANG, J. T.; ZAKI, M. J.; TOIVONEN, H. T.; SHASHA, D. Introduction to data mining in bioinformatics. In: **Data Mining in Bioinformatics**. Springer, 2005. p.3–8.

WANG, S.; MINKU, L. L.; YAO, X. Online class imbalance learning and its applications in fault detection. **International Journal of Computational Intelligence and Applications**, v.12, n.04, p.1340001, 2013.

WANG, S.; YAO, X. Relationships between diversity of classification ensembles and single-class performance measures. **IEEE Transactions on Knowledge and Data Engineering**, v.25, n.1, p.206–219, 2013.

WATSON, J. D.; BAKER, T. A.; BELL, S. P.; GANN, A.; LEVINE, M.; LOSICKE, R. **Biologia molecular do gene**. Artmed Editora, 2015.

WEISS, S.; KULIKOWSKI, C. **Computer systems that learn**. San Mateo, CA: Morgan Kaufmann, 1991.

WITTEN, I.; FRANK, E.; HALL, M. **Data Mining – Practical machine learning tools and techniques**. Burlington, MA: Morgan Kaufmann Publishers, 2011.

WITVLIET, D.; STROKACH, A.; GIRALDO-FORERO, A. F.; TEYRA, J.; COLAK, R.; KIM, P. M. ELASPIC web-server: proteome-wide structure based prediction of mutation effects on protein stability and binding affinity. **Bioinformatics**, p.btw031, 2016.

WORTH, C. L.; BICKERTON, G. R. J.; SCHREYER, A.; FORMAN, J. R.; CHENG, T. M.; LEE, S.; GONG, S.; BURKE, D. F.; BLUNDELL, T. L. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs)

and their relation to disease. **Journal of bioinformatics and computational biology**, v.5, n.06, p.1297–1318, 2007.

WORTH, C. L.; PREISSNER, R.; BLUNDELL, T. L. SDM—a server for predicting effects of mutations on protein stability and malfunction. **Nucleic acids research**, p.gkr363, 2011.

XU, J.; BAASE, W. A.; BALDWIN, E.; MATTHEWS, B. W. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. **Protein Science**, v.7, n.1, p.158–177, 1998.

YANG, P.; YOO, P. D.; FERNANDO, J.; ZHOU, B. B.; ZHANG, Z.; ZOMAYA, A. Y. Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. **IEEE transactions on cybernetics**, v.44, n.3, p.445–455, 2014.

ZAKI, M. J.; MEIRA, W. **Data mining and analysis – fundamental concepts and algorithms**. Cambridge University Press, 2014.

ZHANG, X.; XIAO, W.; ACENCIO, M. L.; LEMKE, N.; WANG, X. An ensemble framework for identifying essential proteins. **BMC bioinformatics**, v.17, n.1, p.322, 2016.

ZHANG, Z.; WANG, L.; GAO, Y.; ZHANG, J.; ZHENIROVSKYY, M.; ALEXOV, E. Predicting folding free energy changes upon single point mutations. **Bioinformatics**, v.28, n.5, p.664–671, 2012.

ZHAO, H.; RAM, S. Entity matching across heterogeneous data sources: An approach based on constrained cascade generalization. **Data & Knowledge Engineering**, v.66, n.3, p.368–381, 2008.

ZHAO, N.; HAN, J. G.; SHYU, C.; KORKIN, D. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. **PLoS Comput Biol**, v.10, n.5, p.e1003592, 2014.

ZHOU, Z. H. **Ensemble methods – foundations and algorithms**. CRC Press, 2012.