MUTAGRAPH: MODELOS E ALGORITMOS PARA PREDIÇÃO NA AFINIDADE DE COMPLEXOS PROTEICOS ATRAVÉS DE GRAPH KERNEL E MÉTRICAS DE REDES COMPLEXAS

LAERTE MATEUS RODRIGUES

MUTAGRAPH: MODELOS E ALGORITMOS PARA PREDIÇÃO NA AFINIDADE DE COMPLEXOS PROTEICOS ATRAVÉS DE GRAPH KERNEL E MÉTRICAS DE REDES COMPLEXAS

Projeto de tese apresentado ao Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: Raquel Cardoso de Melo Minardi
Coorientador: Douglas Eduardo Valente Pires

Belo Horizonte - MG

Outubro de 2017

Agradecimentos

Primeiramente à Deus que permitiu que tudo isso acontecesse, ao longo da minha vida, e não somente nestes anos como doutorando, mas que em todos os momentos é o maior professor que alguém pode conhecer.

Agradeço aos meus pais Maria Eleusa e Raimundo Rodrigues pelo exemplo e motivação e por sempre acreditarem em mim e em especial por serem um exemplo de vida e perseverança.

À minha amada esposa Renata Caroline que me acompanhou desde a aprovação até a defesa. Não tenho palavras para agradecê-la por tal companhia.

Às minhas irmãs Luciene e Lídia pelo carinho e apoio.

Agradeço à minha orientadora Raquel Cardoso de Melo Minardi pela oportunidade a mim concedida, pois uma pergunta feita por ela em nosso primeiro encontro me marcou e vou levar para toda a vida: "Como você espera estar daqui há 10 anos?".

Ao Coorientador Douglas Eduardo Valente Pires pelo tempo dispendido em me ajudar neste trabalho de todas as formas que pôde e mostrando um grande exemplo como professor que levarei durante toda a minha carreira de docente.

Aos meus colegas de laboratório, Wellisson, Pedro, Alexandre e João Arthur pelos bons momentos que compartilhamos no LBS.

Agradeço à FAPEMIG por financiar 3 anos deste trabalho e ao Instituto Federal de Minas Gerais por me liberar de horas de trabalho para a conclusão desta tese.

"O seu foco determina a sua realidade."
(Qui-Gon Jinn – Star Wars Episódio I.)

Resumo

Mutações em regiões codificadoras podem afetar estrutura e função proteicas levando ao seu mau funcionamento, estando esse relacionado a diversas desordens hereditárias além de também estarem relacionadas ao surgimento e predisposição a diversos tipos de cânceres. Mutações do tipo *missense*, onde há troca de um aminoácido por outro, são um tipo comum de alteração genética que pode afetar a função proteica por sua desestabilização e/ou alteração da afinidade entre proteína e seus parceiros, sejam esses pequenas moléculas ou outras proteínas. A despeito de relevantes esforços descritos na literatura com o intuito de elucidar a relação entre mutações *missense* e seu impacto na estabilidade da estrutura proteica e, por conseguinte, sua função, prever o efeito de uma mutação na afinidade de um complexo quaternário proteico ainda é um grande desafio. Interações proteína-proteína são importantes para o desempenho de diversas funções no organismo e são delicadamente reguladas. Entender como mutações afetam a afinidade de complexos proteicos podem auxiliar no entendimento de seu papel em doenças bem como permitir a engenharia de interfaces proteicas para propósitos biotecnológicos. Nesse contexto, apresentamos MutaGraph, uma nova abordagem computacional, quantitativa e baseada na estrutura tridimensional para a predição do efeitos de mutações *missense* na afinidade de complexos proteicos baseada em *graph kernels* e métricas de redes complexas. Utilizando bases de dados que relacionam mutações em interfaces de complexos proteicos com estruturas resolvidas e parâmetros termodinâmicos experimentalmente determinados de seus efeitos, utilizamos técnicas de aprendizado supervisionado para treinar e avaliar modelos preditivos. O MutaGraph conseguiu prever de forma bem sucedida o efeito de mutações em interfaces proteicas, alcançando uma correlação de Pearson de até 0,84 em validação

cruzada. O método proposto está disponível livremente como um servidor web, que implementa técnicas para visualização do efeito de mutações que pode ser acessado em http://bioinfo.umfg.br/mutagraph.

Abstract

Mutations in coding regions can affect the structure and the function of a protein leading to malfunction and still related to hereditary disorders and propensity to several cancers. Missense mutation types, where have the change of one amino acid to another it is a common type of genetic exchange could affect the protein's function by destabilizing and/or affinity change between the protein and others partners, it will be small molecules and other proteins. In spite of relevant efforts described in the literature in elucidating the relationship between the missense mutation and your impact on protein stability structure and therefore your function, predicting your mutation in the affinity of the protein quaternary complex is still a great challenge. Protein-protein interactions are essential for the performance of various functions in the body and are carefully regulated. Understanding how mutations can affect the affinity of protein complexes may aid in understanding their role in diseases as well as providing the engineering of protein interfaces for biotechnological purposes. In this context, we present MutaGraph, a new computational, quantitative and three-dimensional approach based on the prediction of the effects of missense mutations on the affinity of protein complexes based on graph kernels and complex network metrics. Using databases that describe mutations in protein complex interfaces with resolved structures and experimentally determined thermodynamic parameters of their effects, we use supervised learning techniques to train and evaluate predictive models. MutaGraph was able to successfully predict the effect of mutations in protein interfaces, achieving a Pearson correlation of up to 0,84 in cross-validation. The proposed method is freely available as a web server, which implements techniques for visualizing the effect of mutations and can be accessed at http://bioinfo.umfg.br/mutagraph.

Lista de Figuras

1.1 A estrutura geral de todos os 20 aminoácidos juntamente com suas cadeias laterais e sua classificação [Nelson & Cox, 2008]. 2

1.2 Ligações peptídicas entre aminoácidos, os ângulos ϕ e ψ correspondem às ligações ligações N-Cα e C-Cα respectivamente enquanto a ligação ω diz respeito à própria ligação que ocorre entre o grupo carboxílico de aminoácido com o grupo amino de outro [Nelson & Cox, 2008]. 3

1.3 Exemplo do processo de interação das subunidades da Hemoglobina. Quando há ausência de ligante esta região é flexível facilitando sua ligação e quando isto ocorre consequentemente temos uma estabilização na estrutura na região que antes menos estável e a interação entre os complexos também aumenta. 4

1.4 Codificação dos tripletos para a expressão de proteínas [Nelson & Cox, 2008]. 5

1.5 Representação de grafos. Em (a) é apresentado um grafo não direcionado enquanto que em (b) as arestas do grafo G' são direcionadas. . . . 9

1.6 Exemplo do processo de regressão. 12

1.7 Interface (cor amarelo) das cadeias A e B (cor azul e verde respectivamente) do PDB 2YPI (Triose-fosfato isomerase) sendo a cadeia A da cor cinza e a cadeia B da cor verde. 14

3.1 Esquema do processamento do modelo. As etapas de construção da estrutura mutante e selvagem consistem em gerar a estrutura normalizada do Selvagem (adicionar átomos faltantes) e construir a mutante pelo software FoldX. Já a etapa de extração de *features* computa todos os atributos utilizados pelos algoritmos de regressão e para visualização. A etapa de treino e teste farão a predição da mutação baseado nas informações contidas na base de dados SKEMPI, e, por fim, a visualização disponibiliza ao usuário uma interface para interagir com o resultado das etapas anteriores. 22
3.2 Histograma dos valores de energia livre de ligação das mutações extraídas da base de dados SKEMPI utilizadas neste trabalho. 25
3.3 Representação da interface a ser analisada. As cadeias E, F e G estão coloridas de azul enquanto a cadeia I de cor verde. Assim, a interface considerada para análise é a feita entre a cadeia I com qualquer uma das outras 3. 27
3.4 Representação da distância geodésica e coeficiente de agrupamento (ambos normalizados) das 2.088 estruturas utilizadas neste trabalho (2.007 estruturas mutantes mais 81 estruturas) juntamente com a relação de exemplo de uma rede Aleatória e outra Regular para fins de referência. 30
3.5 Representação do processo de produto direto entre 2 grafos. Basicamente é a aplicação do produto direto em ambas as matrizes de adjacência. 35
3.6 Decomposição de 2 imagens distintas com o algarismo 2. No final do processo de decomposição de ambas as imagens estas ficaram mais semelhantes entre si do que as originais. 40
3.7 Representação gráfica do coeficiente de Pearson para três casos distintos. Em (a) temos uma correlação direta, em (b) temos uma correlação indireta e em (c) quando não há correlação linear. 42
3.8 Representação do modelo de regressão por consenso para quatro diferentes algoritmos (Processo Gaussiano, Árvores M5, KNN e SVR) nos quais o resultado destes formarão o conjunto de atributos da árvore de regressão. 43

3.9 Árvore de regras M5 (a) para um conjunto de dados num espaço bidimensional tendo uma terceira dimensão como a variável de interesse (b) [Rahimikhoob et al., 2013]. 44
3.10 Execução do algoritmo KNN. O losango representa a instância de interesse enquanto os círculos vermelhos as instâncias do conjunto de Treino. Em 3.10b temos a instância com a variável de interesse é predita pela média dos 3 instâncias mais próximas. 45
3.11 Construção do modelo de regressão criado pelo SVR mostrando o uso das variáveis de erro (ϵ) e de custo (ζ) para a construção da margem de perda [Schölkopf et al., 1998]. 46
3.12 Esquema de construção de floresta de *Random Forest Tree*. O conjunto de dados é separado em diferentes conjuntos de amostras afim de gerar um conjunto de árvore. Por fim o consenso entre o conjunto de cada uma delas dará o resultado final do modelo. 47

4.1 Correlação de Pearson fazendo a regressão com o algoritmo KNN, o valor de K variando entre 1 e 25. 50
4.2 Progressão da correlação de Pearson pela quantidade de atributos removidos para cada um dos algoritmos de aprendizado supervisionado. . 52
4.3 Box Plot da correlação de Pearson para as 10 dobras na validação cruzada de cada um dos algoritmos utilizados. 54
4.4 Box Plot da correlação de Pearson para as 10 dobras na validação cruzada de cada um dos algoritmos utilizados usando o SVD para redução de dimensionalidade e remoção de ruídos. 56
4.5 Curva de valores singulares de cada um dos algoritmos de regressão. . . 57
4.6 Correlação de Pearson para o conjunto de mutações T55 utilizando o modelo construído apenas com o conjunto de mutações com $rho = -0.031$. 60
4.7 Correlação de Pearson para o conjunto de mutações T56 utilizando o modelo construído apenas com o conjunto de mutações $rho = 0.006$. . . 61

5.1 Página principal do *webservice*, a partir dela é possível acessar as funções de criar um novo job e consultar o status da que já foram criadas. 63

5.2 Tela para seleção do conjunto de mutações que será processada. Nesta etapa o usuário é capaz de selecionar a cadeia, posição e o resíduo mutante. 64
5.3 Ação de confirmação da remoção de uma determinada mutação da *Job*. 65
5.4 Etapa de seleção das cadeias na ferramenta. Todas as cadeias do grupo 1 são destacadas com a cor vermelha enquanto o grupo 2 com a cor verde. Quando uma cadeia não faz parte do grupo então fica com a cor cinza. 66
5.5 Página de conclusão da criação da Job. O usuário pode consultar pelo código o status da *Job* ou também pode registrar seu e-mail para ser notificado quando o processamento terminar. 67
5.6 Texto com a conclusão do processo de construção e análise da mutação. A partir do próprio e-mail o usuário é capaz de acessar o *webservice* na página de cada Job que requisitou. 67
5.7 Página com as informações do efeito da mutação onde (a) é a nomenclatura utilizada para a mutação, (b) o valor da variação de energia livre de ligação predita pela ferramenta, (c) a estrutura de visualização do grafo de contatos, (d) a estrutura da proteína do mutante e selvagem sobrepostos e (d) o menu de contexto. 68
5.8 Legenda da ferramenta contendo a descrição das cores, formas das arestas e e vértices além das possíveis ações existentes para interação. . . . 69

Lista de Tabelas

3.1 Relação de estruturas depositadas no PDB com incosistências. 24
3.2 Relação dos atributos extraídos das estruturas. 38

4.1 Correlação de Pearson juntamente com o desvio padrão (em $KCal/mol^{-1}$) para cada um dos grupos de atributos nos 4 algoritmos de regressão utilizados neste trabalho. 51
4.2 Quantidade de atributos utilizados em cada algoritmo e seu coeficiente de Pearson para a predição da afinidade do complexo proteico. 53
4.3 Tabela com os resultados referentes ao uso do SVD apresentando o ganho no coeficiente de Pearson em relação à Tabela 4.2 juntamente com o teste de significância deste ganho desta correlação. 55
4.4 Comparação do MutaGraph com os softwares BeAtMuSiC e mCSM nos casos de validação cruzada de 10 dobras,não redundante em posição e proteína. 58
4.5 Comparação do MutaGraph com os software MutaBind no cenários de validação cruzada de 5 e 4 além do conjunto de mutações apenas com mutações em proteínas inibidoras de protease em dobras com complexos proteícos similares (SKEMPI_MutaBind). 59
4.6 *Kendall's score* do MutaGraph com demais trabalhos no conjunto T55. 60
4.7 *Kendall's score* do mutagraph com demais trabalhos no conjunto T56. 61

Sumário

Agradecimentos **v**

Resumo **ix**

Abstract **xi**

Lista de Figuras **xiii**

Lista de Tabelas **xvii**

1 Introdução **1**

1.1 Objetivo geral 6

1.2 Objetivos específicos 7

1.3 Justificativa 7

1.4 Fundamentação teórica 8

1.4.1 Grafos 8

1.4.2 Redes complexas 8

1.4.3 Aprendizado Supervisionado 11

1.4.4 Regressão 12

1.4.5 Interação Proteína-Proteína (PPI[1]) 13

1.5 Estrutura do trabalho 15

2 Revisão de literatura **17**

2.1 Base de dados de mutações 17

[1]Do inglês *protein-protein Interaction*

2.2 Ferramentas de predição do efeito de mutações em interações proteína-proteína 18

3 Materiais e Métodos **21**

3.1 Conjunto de dados SKEMPI 23
3.2 Construção das estruturas mutantes 25
3.3 Variação da energia livre de ligação $\Delta\Delta G_{bind}$ 26
3.4 Construção da rede de contatos 26
3.5 Caracterização da rede de contatos 28
3.6 Extração de *features* 29
3.6.1 Métricas de topologia de rede 29
3.6.2 Graph kernel 34
3.6.3 Atributos auxiliares 36
3.6.4 Assinatura da mutação 37
3.7 Redução de dimensionalidade 39
3.8 Avaliação do modelo 39
3.8.1 Validação cruzada de baixa redundância 41
3.9 Métricas de avaliação 41
3.9.1 Coeficiente de correlação de Pearson 41
3.10 Teste estatístico para significância de diferenças de correlação . . . 41
3.11 Processo de regressão 42
3.11.1 Processo Gaussiano 43
3.11.2 Árvore de Regras M5 43
3.11.3 k-Nearest Neighbors 45
3.11.4 Vetores de suporte de Regressão (SVR[2]) 45
3.11.5 *Random Forest Tree* 47

4 Resultados e discussões **49**

4.1 Construção do modelo de Regressão 49
4.1.1 Seleção do K para o algoritmo KNN 49
4.1.2 Caracterização dos atributos utilizados 51
4.1.3 Seleção de atributos 52

[2]Do inglês *Support Vector Regression*

4.1.4 Redução de dimensionalidade 54
4.1.5 Modelo de Regressão final 55
4.2 Comparação com outros trabalhos da literatura 58
4.2.1 BeAtMuSiC e mCSM 58
4.2.2 MutaBind 59
4.3 Estudo de Caso: Conjunto de mutações do CAPRI (Rodada 26) . . 59

5 Web Service **63**
5.1 Construção do *Job* 64
5.2 Visualização do efeito da mutação 67

6 Conclusões e trabalhos futuros **71**
6.1 Trabalhos futuros 72

Anexo A Resultado da predição **73**

Anexo B Atributos utilizados no processo de regressão **101**

Referências Bibliográficas **105**

Capítulo 1

Introdução

As proteínas são os compostos orgânicos mais abundantes em organismos vivos, sendo encontradas em todas as partes da célula e fundamentais em todos os aspectos da estrutura e função celular [Nelson & Cox, 2008]. De fato, proteínas são macromoléculas versáteis formadas por cadeias de aminoácidos de grande importância para a manutenção da vida e seu mau funcionamento, por sua vez, está relacionado com diversas desordens hereditárias, bem como predisposição e surgimento de diversos tipos de tumores [Klug et al., 2009].

As proteínas são compostas principalmente por átomos de carbono, nitrogênio, oxigênio e enxofre. Algumas proteínas contêm elementos adicionais, particularmente fósforo, ferro, zinco e cobre. Existem 20 tipos de aminoácidos comumente encontrados nos seres vivos e cada um deles apresenta uma estrutura em comum onde um Carbono (chamado carbono alfa) liga-se a um grupo amino, um grupo carboxílico e a uma cadeia lateral com dimensões e características variáveis. A Figura 1.1 apresenta a diversidade de cadeias laterais dentre os aminoácidos naturais e suas propriedades físico-químicas.

Proteínas são polímeros formados por cadeias polipeptídicas compostas por resíduos de aminoácidos unidos por ligações peptídicas. A sequência desses aminoácidos compõe a estrutura primária proteica.

A cadeia principal proteica possui certa flexibilidade conferida pela rotação de suas ligações formando os ângulos diedrais *phi* (ϕ), *psi* (ψ) e *ômega* (ω) no qual são encontrados nas ligações entre os Carbonos alfa de aminoácidos vizinhos,

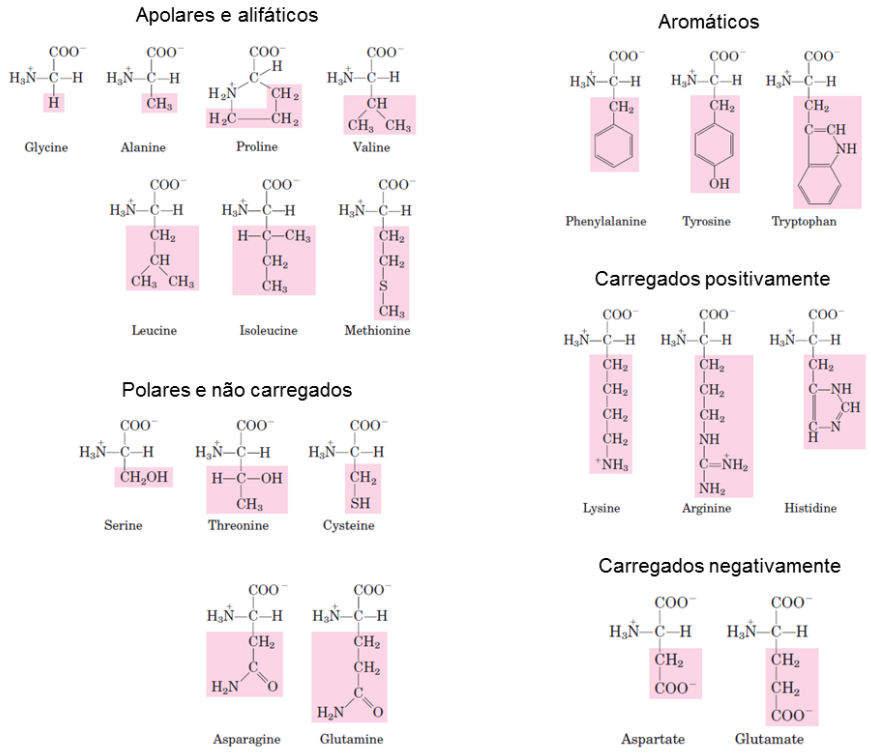


Figura 1.1: A estrutura geral de todos os 20 aminoácidos juntamente com suas cadeias laterais e sua classificação [Nelson & Cox, 2008].

a Figura 1.2 apresenta a posição espacial de cada um destes ângulos numa cadeia polipeptídica. O ângulo ômega (ω) corresponde à torção da ligação peptídica, sendo na maioria dos casos trans (ângulo de 180°) e em casos mais específicos, como por exemplo a prolina, onde este ângulo assume 0° (ou seja, cis). Já os ângulos *phi* e *psi* correspondem às ligações N-Cα e C-Cα respectivamente, entretanto nem todos os valores para estes ângulos são aceitos por que causariam choques estéricos [Berg et al., 2014].

Por meio de torções desses ângulos as estruturas secundárias são formadas, como as hélices (sendo as alfa-hélices as mais comuns), fitas (que por sua vez formam folhas-beta paralelas e antiparalelas) e voltas que basicamente descrevem o arranjo espacial da cadeia principal da proteína. Estas conformações resultam tanto dos ângulos assumidos por ϕ, ψ e ω quanto as ligações de hidrogênio que as cadeias laterais formam [Berg et al., 2014].

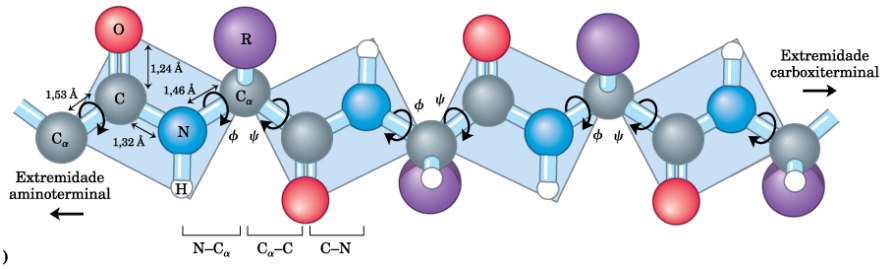


Figura 1.2: Ligações peptídicas entre aminoácidos, os ângulos ϕ e ψ correspondem às ligações ligações N-Cα e C-Cα respectivamente enquanto a ligação ω diz respeito à própria ligação que ocorre entre o grupo carboxílico de aminoácido com o grupo amino de outro [Nelson & Cox, 2008].

Essas estruturas organizam-se, então, em uma estrutura tridimensional. A disposição espacial de uma proteína cria uma conformação tridimensional específica e estável, responsável pela funcionalidade da proteína conhecida por estrutura terciária ou nativa. A estabilidade da estrutura terciária proteica, bem como a sua conformação tridimensional é determinada por uma série de forças e interações, como ligações dissulfeto (covalentes), pontes salinas também conhecidas como interações eletrostáticas, ligações de hidrogênio, interações hidrofóbicas. Além das forças de atração, há ainda as forças de repulsão, principalmente eletrostáticas, que são muito importantes no balanço energético para a estabilização da proteína [Vagenende et al., 2009].

Determinadas unidades biológicas podem ser formadas pela união de 2 ou mais estruturas terciárias (cadeias) que podem ser diferentes ou idênticas que se conectam por meio de interações não covalentes, formando o que chamamos de estruturas quaternárias. Muitas interações entre proteínas (do inglês *Protein-Protein Interaction*) são formados de forma permanente ou transiente para execução de determinada função biológica de modo que sua interação é delicadamente regulada tanto por subunidades idênticas que interagem entre si (homo-oligômero) ou de subunidades distintas (hétero-oligômeros) [Jones & Thornton, 1996, Zhang et al., 2017].

Um fato importante na questão interação entre complexos proteicos está na região em que a interação ocorre. Levando em conta que a superfície das proteínas é constituída em sua maioria por resíduos polares, e, consequentemente a forma-

ção de interfaces nos complexos proteicos dá-se pelo enterro de resíduos polares e carregados é diferente da interação com outros tipos de complexos, como por exemplo peptídeos pequenos em regiões específicas na superfície onde a interação ocorrerá numa cavidade em específico alterando a conformação da cavidade.

Como exemplo temos a Hemoglobina na qual as subunidades possuem uma interação entre si de forma que quando o ligante (neste caso o oxigênio) liga com uma subunidade então a interação no complexo é alterada aumentando sua afinidade. A Figura 1.3 apresenta o processo a mudança de interação do complexo proteico entre si quanto com o ligante.

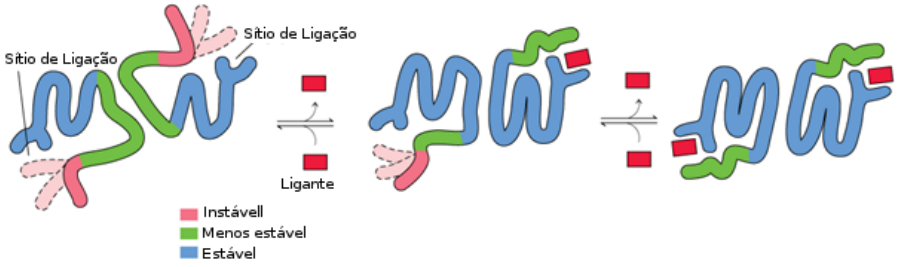


Figura 1.3: Exemplo do processo de interação das subunidades da Hemoglobina. Quando há ausência de ligante esta região é flexível facilitando sua ligação e quando isto ocorre consequentemente temos uma estabilização na estrutura na região que antes menos estável e a interação entre os complexos também aumenta.

Para a manutenção da vida, diversos complexos proteicos possuem funções específicas expressas pelo código genético. A modificação de um único nucleotídeo na sequência do DNA é considerado uma mutação sendo que esta alteração pode ocorrer tanto por erro nas vias de reparo do DNA como por pressão evolutiva. Apesar disso, esta substituição não necessariamente pode alterar a proteína codificada pois o código genético é degenerado. Como apresentado na Figura 1.4 a Prolina, por exemplo, é sintetizada quando os 2 primeiros códons são citosinas (uma vez que o terceiro códon se for alterado não influenciará na proteína expressa), enquanto o ácido aspártico é codificado pelas sequências GAU e GAC, se, alterado o último nucleotídeo da sequência para uma adenina o resultado será o ácido glutâmico fazendo com que a proteína expressa seja diferente.

A substituição de um único nucleotídeo na região codificadora é chamado de SNP (*Single Nucleotide Polymorphism*) que pode causar um polimorfismo sinô-

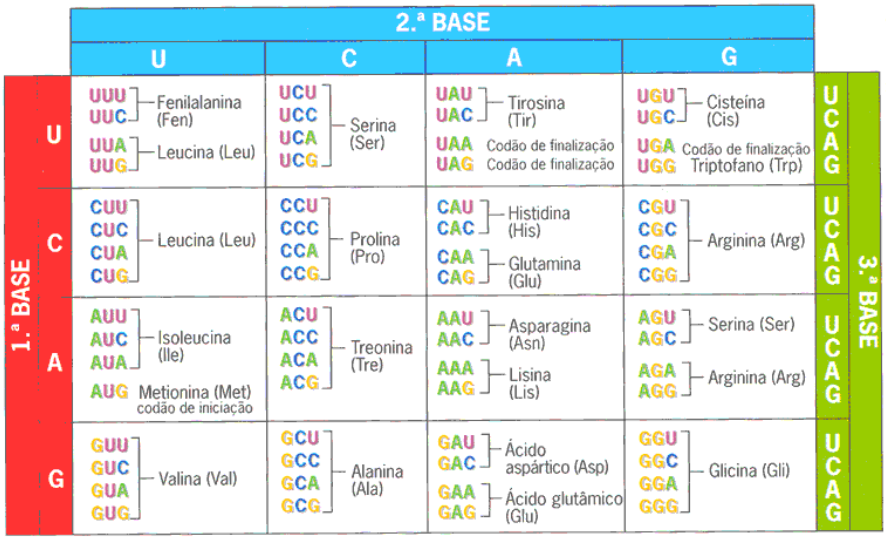


Figura 1.4: Codificação dos tripletos para a expressão de proteínas [Nelson & Cox, 2008].

nimo (sSNP – *synonymous SNP*), onde o nucleotídeo substituído apenas muda o códon mas não o aminoácido que será sintetizado. Já o polimorfismo não-sinônimo (nsSNP - *non-synonymous SNP*) além de modificar o códon também modifica o aminoácido que será traduzido. A alteração do nucleotídeo pode não somente mudar o aminoácido expresso mas também pode ser modificado para um códon de parada (sendo esse tipo de efeito chamado de mutação *nonsense*) onde a expressão da proteína é parada de forma prematura. Mas neste trabalho o objetivo é avaliar mutações de natureza *missense*, onde há troca do aminoácido codificado por outro. Um dos casos mais conhecidos de nsSNP é a mutação responsável pela anemia falciforme em que a hemoglobina muda sua forma bicôncava para um formato similar a de uma foice. Esta modificação estrutural deve-se a substituição de uma timina por adenina no gene que codifica a hemoglobina que ocasiona na substituição do Ácido glutâmico da posição 6 da cadeia β por uma Valina [Galiza Neto & Pitombeira, 2003].

Um outro impacto que pode ocorrer em uma proteína que sofreu uma mutação é a alteração da afinidade entre as cadeias do complexo proteico. A interação entre as suas subunidades pode ser alterada de forma que aumente, diminua ou mantenha a afinidade entre elas.

Entender o funcionamento bioquímico de uma mutação é considerado o primeiro passo para decifrar o vínculo da variação genética com determinadas doenças. Ainda, compreender as redes de interações estabelecidas por estas associadas a doenças é relevante para o estudo de desordens genéticas complexas, como câncer, autismo e diabetes [Bergholdt et al., 2012, O'Roak et al., 2012, Wu et al., 2010, Zhao et al., 2014].

Diferentes métodos já foram propostos para predizer o efeito de mutações que afetam a termoestabilidade utilizando informações extraídas da sequência da proteína quanto de sua estrutura ou ambas e da mesma forma para o problema de identificar como mutações afetam a afinidade de complexos proteicos. Ainda há muito que ser explorado sendo um desafio mais recente [Pires et al., 2014, Dehouck et al., 2013, Li et al., 2016].

Sendo assim, o objetivo deste trabalho é contribuir de forma a propor um novo modelo computacional capaz de predizer o efeito de uma mutação na afinidade proteína-proteína utilizando métricas de redes, *graph kernel* e informações estruturais.

Diferentes aplicações de grafos para resolver problemas biológicos existem na literatura como por exemplo a modelagem de redes de vias metabólicas [Zhang & Wiemann, 2009], modelagem de interação proteína-proteína, regulação de genes [Aittokallio & Schwikowski, 2006] e ainda para predição de flexibilidade em estruturas proteicas [Jacobs et al., 2001]. O presente trabalho visa, pela primeira vez, utilizar tanto a modelagem dos contatos existentes na interface quanto o uso de atributos de redes complexas e *graph kernel* como método para construir um modelo de predição do efeito de mutações em complexos proteicos.

1.1 Objetivo geral

O principal objetivo deste trabalho é propor, avaliar e validar uma metodologia que utilize métricas de redes complexas e *graph kernel* juntamente com informações da estrutura da proteína afim de predizer o efeito de uma mutação *missense* na afinidade de complexos proteína-proteína.

1.2 Objetivos específicos

1. Extrair as mutações SNP *missense* da base de mutações que afetam o complexo proteico SKEMPI;

2. Analisar a correlação entre o efeito de mutações em interfaces proteína-proteína e atributos estruturais e de sequência e identificar atributos preditivos;

3. Propor e avaliar uma modelagem de mutações por meio de teoria de grafos e avaliar a contribuição de métricas de redes complexas e *graph kernel* como atributo preditivo;

4. Construir uma interface *web* amigável para o melhor modelo treinado;

5. Analisar a semântica das predições obtidas pelo modelo;

1.3 Justificativa

Nos últimos anos os conjuntos de dados experimentais referentes as mutações que afetam a função da proteína vêm aumentando consideravelmente principalmente para mutações que afetam a termoestabilidade. Com estes dados foi possível criar diferentes métodos computacionais para prever o efeito de uma mutação na termoestabilidade de uma proteína.

Entretanto, pouco que se fez na área para mutações que afetam a afinidade proteína-proteína. Elucidar e prever o efeito desse tipo de mutação ainda é um desafio recente e em aberto uma vez que mutações que afetam o complexo proteico podem ser responsáveis por doenças [Bullock et al., 2000, O'Roak et al., 2012].

Atualmente existe uma demanda para a predição da variação da afinidade de proteína uma vez que existem diversas implicações para o seu uso como por exemplo no processo de engenharia de proteínas para a criação ou alteração de proteínas com níveis específicos de estabilidade, atividade enzimática e/ou com a ligação com outras moléculas [Potapov et al., 2009].

Mutações que afetam a afinidade entre complexos relacionadas a doenças podem estar ligadas a mudanças em vias metabólicas, e, com a ajuda de métodos

computacionais podem desempenhar um papel fundamental para o entendimento e reconstrução destas vias [Wu et al., 2010].

1.4 Fundamentação teórica

Para um melhor entendimento do problema e da solução proposta abaixo serão apresentados conceitos teóricos utilizados durante o texto.

1.4.1 Grafos

O grafo é um modelo matemático que tem por objetivo representar o relacionamento entre diferentes elementos de um conjunto. Relacionamentos entre pessoas em redes sociais, links entre sites e diversos outros conjuntos podem ser modelados por grafos uma vez que cada entidade pode estar conectada ou não com outros.

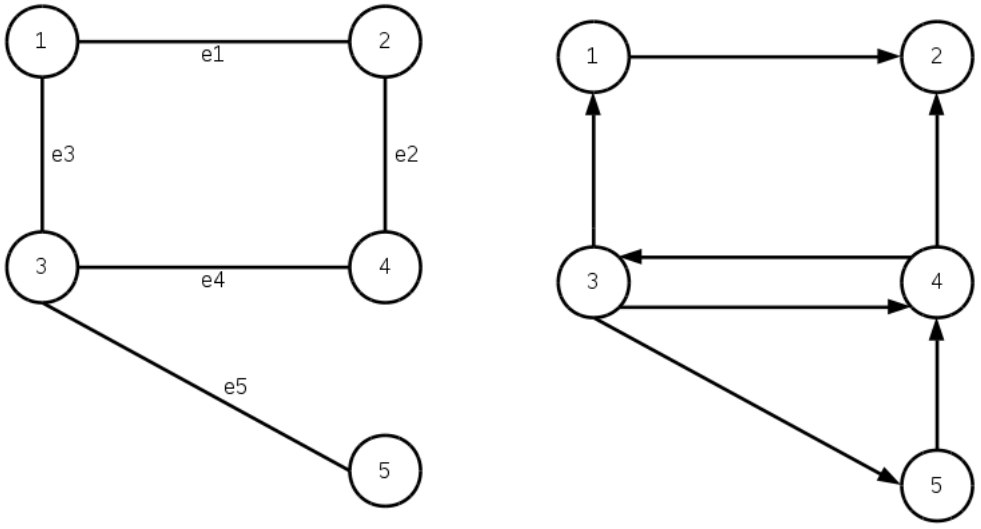
Matematicamente podemos considerar um grafo como pela representação $G = (V, E)$ onde V representa o conjunto de vértices e E o conjunto de arestas onde uma aresta e é denotada por $e = (V_i, V_j)$, ou seja, dizemos que dois vértices (neste caso V_i e V_j) estão conectados entre si. A Figura 1.5 mostra a representação visual do grafo G com vértices $V = \{1, 2, 3, 4, 5\}$ e as arestas $E = \{e1, e2, e3, e4, e5\}$ que ligam os vértices.

Também podemos construir grafos nos quais as arestas possuem direção, chamados de grafos direcionados; as arestas possuem um vértice de origem e outra de destino tornando $(V_i, V_j) \neq (V_j, V_i)$. Visualmente a aresta passa a ser representada como uma seta. A Figura 1.5b apresenta visualmente um grafo direcionado G', nele as setas indicam a direção da aresta.

Podemos ainda definir o peso das arestas de um grafo denotado por $w(V_i, V_j)$ onde w indica o peso da aresta entre V_i e V_j [Golumbic, 2004] sendo este denominado de grafo ponderado ou rotulado.

1.4.2 Redes complexas

Um grafo que modela o relacionamento entre entidades é considerado uma rede complexa e que exista um grande número de unidades interconectas entre si. Como



(a) Representação visual do grafo G (b) Representação visual do grafo G'

Figura 1.5: Representação de grafos. Em (a) é apresentado um grafo não direcionado enquanto que em (b) as arestas do grafo G' são direcionadas.

exemplo, podemos modelar o sistema de transporte aéreo como um grafo onde os vértices representam as cidades e as arestas os voos entre elas. Em outras palavras uma rede complexa tem por objetivo auxiliar na compreensão do relacionamento entre os objetos de alvo do estudo.

Definir as propriedades que irão representar tanto os vértices os quais constituem como unidades básicas do nosso modelo e os vértices, que por sua vez representam a interação entre estes elementos é um passo crucial para a sua construção da rede, pois, estas informações irão definir a sua topologia [Boccaletti et al., 2006].

Sendo assim, ao criarmos uma rede complexa temos o objetivo de extrair informações relevantes ali encontradas, como por exemplo fenômenos ou características que se destacam em função do relacionamento dos objetos. As propriedades estruturais, ou características topológicas nos dá o suporte necessário para o entendimento das informações ali encontradas [Girvan & Newman, 2002, Broder et al., 2000].

Redes complexas também podem ser classificadas segundo o seu tipo estrutural como apresentado a seguir:

1. **Rede regular**: Todos os nós possuem o mesma quantidade de arestas.

2. **Redes aleatórias**: A conexão entre 2 vértices pode ocorrer segundo uma probabilidade p sendo que a quantidade de arestas criadas segue uma distribuição de Poisson[1]. Desta forma, o grau[2] esperado de um vértice qualquer nesta rede pode ser definido pela Equação 1.1.

$$\langle k \rangle = p(N-1) \tag{1.1}$$

   Onde N, p e $\langle k \rangle$ representam a quantidade de vértices da rede, a probabilidade do vértice se conectar a outro e o grau médio de cada vértice respectivamente [Erdös & Rényi, 1970].

3. **Redes de pequeno mundo (*small-world*)**: Segundo [Watts & Strogatz, 1998] nesta estrutura é possível alcançar qualquer um dos vértices a partir de outros em poucos passos devido à proximidade entre eles. Este tipo de estrutura pode ser utilizado para representar conexões entre sites da internet, identificação de resíduos importantes para interações proteína-proteína e na organização e funcionamento de células [Barabasi & Oltvai, 2004, Adamic, 1999, Vendruscolo et al., 2002].

   Uma característica básica de uma rede de pequeno mundo está em sua baixa distância geodésica e baixo grau de agrupamento. Em outras palavras, dado dois vértices quaisquer existem alta probabilidade de existirem vizinhos em comum entre si. Espera-se que o caminho mínimo entre dois vértices escolhidos espera-se a aproximação de $L(v_i, v_j) \approx ln(|V|)$ onde $|V|$ é a quantidade total de vértices do grafo [Bassett & Bullmore, 2016, Hiromoto, 2016, Zhang et al., 2013].

4. **Redes livre de escala**: A principal característica deste modelo está em seu crescimento, onde novos vértices e arestas são adicionados incrementalmente. Ao inserir um novo vértice nesta rede existe a tendência de vértices com de

[1] É uma distribuição discreta de probabilidade aplicada a ocorrências de um evento em um intervalo especificado previamente.

[2] Quantidade de conexões que um determinado nó possui na rede, mais detalhes na Seção 3.6.1.

maior grau possuírem uma probabilidade maior de se conectarem com este novo vértice, assim, a natureza desta rede tende a ter poucos nós com muitas conexões e muitos nós com poucas conexões. A probabilidade de um nó se conectar à outro segue a distribuição da lei de potência apresentada na Equação 1.2 onde k é o número de arestas e γ o expoente da lei de potência, assim, é esperado que existam *hubs*[3] neste modelo [Barabási & Albert, 1999].

$$P(x) = k^{-\gamma} \tag{1.2}$$

1.4.3 Aprendizado Supervisionado

O método de aprendizado supervisionado dá-se por algoritmos que recebem como entrada um conjunto de dados já rotulado com a resposta esperado e se chama de conjunto de treino. Os algoritmos de aprendizado supervisionado constroem seu modelo de predição baseado nas cateterísticas observadas no conjunto de treino. A partir do modelo criado o algoritmo é capaz de identificar o rótulo de uma instância que não existe na base de conjunto de treino [Michalski et al., 2013].

Chamados por método de aprendizado supervisionado pois o conjunto de treino pode ser considerado um professor. A partir disto, o algoritmo iterativamente irá gerar o rótulo de cada instância do conjunto de treino e comparar com o que foi informado pelo conjunto de teste de modo a avaliar a qualidade da predição e por conseguinte irá ajustar o modelo para que o resultado se aproxime o máximo possível do rótulo de treino.

O rótulo do conjunto de treino pode ser tanto discreto quanto contínuo. Uma tarefa de classificação por exemplo é a identificação de qual o rótulo (nesse caso, a classe) a instância pertence enquanto rótulos com valores contínuos constroem modelos de regressão [Stuart & Peter, 2016].

A avaliação do modelo é feito pelo método de validação cruzada, que basicamente consiste na construção de cenário no qual uma parte do conjunto de treino é utilizado para o próprio treino enquanto uma outra parte é utilizada como teste afim de simular uma situação real (de prever uma instância não conhecida pelo

[3] Alguns nós de uma rede com grande quantidade de arestas ligadas a ele.

algoritmo). As Seções 3.8 e 3.9 apresentam as métricas e as técnicas de avaliação do modelo de aprendizado supervisionado.

1.4.4 Regressão

O processo de regressão tem por objetivo inferir uma relação entre uma variável de resposta com uma ou várias variáveis explicatórias formando uma equação. Dado um conjunto de n pontos de $D = \{x_i, y_i\}_{i=1}^{n}$ onde x representa a variável explicatória e y a variável de resposta e temos o objetivo de construir uma função $f(x)$ que represente a correlação $y_i = f(x_i) + \epsilon$ de forma que ϵ tenha uma distribuição normal.

A análise de regressão consiste em encontrar uma correlação razoável entre a(s) variável(is) explicatória(s) e a variável de resposta como uma função que pode ser linear ou não. De posse da função $f(x)$ podemos inferir um valor $(\hat{y})$ para uma entrada ainda não analisada $(\hat{x})$. Como exemplo, a Figura 1.6 apresenta o modelo de construa de uma função baseada em um conjunto de pontos e a predição de um novo valor baseado na função construída [Bates & Watts, 2008].

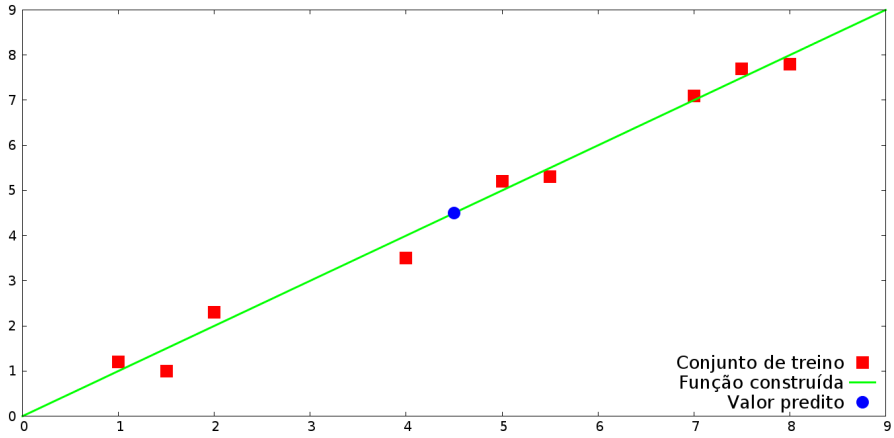


Figura 1.6: Exemplo do processo de regressão.

Levando em conta que o objetivo deste trabalho está na construção de um modelo de predição, a qualidade do resultado predito será avaliado por uma correlação linear entre o resultado obtido pelo modelo em relação à uma amostra real. Para isso será utilizado correlação de Pearson descrita na Seção 3.9.1 mas

também existe a correlação de Spearman na qual avalia a monotonicidade da correlação entre as variáveis (ou seja, não precisa ser uma relação direta entre elas) [Hauke & Kossowski, 2011, Puth et al., 2015].

1.4.5 Interação Proteína-Proteína (PPI[4])

Como já apresentado anteriormente, proteínas podem interagir com outras formando a estrutura quaternária sendo que este fenômeno está ligado ao desempenho da função proteica, sendo essa interação conhecida como interação proteína-proteína. Em organismos procariotos a maioria das estruturas possuem apenas um único domínio[5] [Ekman et al., 2005] e em seres eucariontes multicelulares aparecem proteínas com múltiplos domínios [Li et al., 2001]. Essas interações podem ocorrer de forma que a função da proteína exista se e somente se exista a associação, bem como a associação/desassociação ocorra continuamente, de forma que a funcionalidade exista de forma e estritamente regulada [Fornili et al., 2013].

As forças que regem as interações entre as cadeias de um complexo podem ser efeitos hidrofóbicos, pontes de hidrogênio, forças eletrostáticas dentre outros [Metz et al., 2011, Mazar, 2008, Engin et al., 2012]. A energia livre de Gibbs da formação do complexo (ou energia livre de ligação) pode ser mensurada a partir do equilíbrio da reação (Representadas por K_d e K_a que representam as constantes de associação e desassociação respectivamente sendo calculadas a partir da concentração do complexo livre - interação - com o complexo formado num estado de equilíbrio termodinâmico) sendo passível mensurar quão estável é a interação [Keskin et al., 2008].

Atualmente as técnicas que a literatura apresenta para definir a interação entre os complexos de forma experimental é o método binário (que verifica a interação física entre pares de complexos) ou por co-complexo, no qual verifica se existe interação entre as subunidades sem distinguir interações por pares [De Las Rivas & Fontanillo, 2010]. Em contrapartida, tais métodos são limitados em relação ao custo e escalabilidade comparados aos métodos computacionais para definir se existe ou não estas interações.

[4]Do inglês *protein-protein Interaction*

[5]Parte da cadeia polipeptídica que pode se enovelar de forma independente, formando uma estrutura estável.

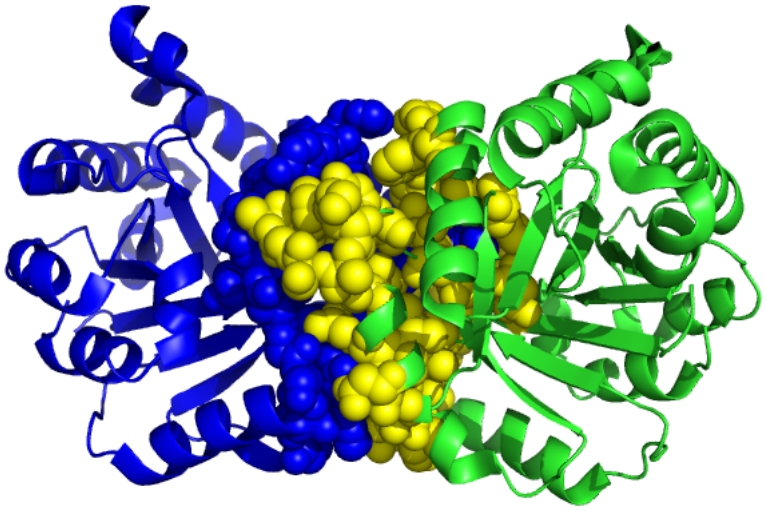


Figura 1.7: Interface (cor amarelo) das cadeias A e B (cor azul e verde respectivamente) do PDB 2YPI (Triose-fosfato isomerase) sendo a cadeia A da cor cinza e a cadeia B da cor verde.

As interações são feitas na interface, que é uma região da cadeia polipeptídica onde os resíduos de ambas as estruturas. É possível caracterizar a interface a partir da sua área (em $Å^2$), complementariedade entre as subunidades, hidrofobicidade dos resíduos, conformação do complexo dentre outros [De Las Rivas & Fontanillo, 2010, Qi et al., 2006]. A Figura 1.7 apresenta a interação[6] entre as cadeias A e B do PDB 2YPI (Triose-fosfato isomerase).

O impacto da mutação num complexo proteico pode causar tanto o aumento quanto a diminuição da afinidade o complexo. Em termo práticos as interações não covalentes entres os complexos podem ficar mais forte ou mais fraco.

[6]Esta interface foi calculada baseando-se na perda da acessibilidade ao solvente dos resíduos quando comparados apenas com seus monômeros.

1.5 Estrutura do trabalho

Este trabalho está organizado da seguinte forma: O próximo Capítulo apresenta a revisão de literatura de trabalhos correlatos ao trabalho. No terceiro Capítulo os materiais e os métodos utilizados para a predição do impacto da mutação. O Capítulo 4 apresenta os resultados obtidos no modelo de predição e ainda a comparação com outros trabalhos da literatura enquanto o capítulo 5 apresenta o *webservice* desenvolvido para ajudar na avaliação da semântica do resultado, e, por último capítulo apresenta as conclusões e os trabalhos futuros e perspectivas.

Capítulo 2

Revisão de literatura

O objetivo deste Capítulo é apresentar os principais trabalhos da literatura recente em relação ao estudo de mutações em interfaces proteína-proteína, bancos de dados e métodos preditivos existentes na literatura até o momento.

2.1 Base de dados de mutações

Existem diferentes bases de dados referentes a mutações, mas a base de dados SKEMPI [1] [Moal & Fernández-Recio, 2012] é específica para mutações que alteram a afinidade de complexos proteicos.

Esta base de dados atualmente possui 3.047 mutações registradas de 85 complexos proteicos diferentes sendo que todas estas mutações são descritas na literatura. Para cada mutação informações como temperatura, afinidade (em molar) e o código do artigo em que o mesmo se encontram podem ser recuperados desta base.

Com estas informações é possível criar modelo computacional capaz de avaliar o efeito de uma mutação no complexo proteico em larga escala.

[1] *Structural database of Kinetics and Energetics of Mutant Protein Interactions*

2.2 Ferramentas de predição do efeito de mutações em interações proteína-proteína

[Zhao et al., 2014] desenvolveu a SNPIN (non-synonymous SNP INteraction). Esta ferramenta usa o método semi-supervisionado afim de predizer o efeito de uma mutação no complexo proteico. O trabalho apresenta 2 tipos diferentes de respostas pelo modelo, o primeiro é a classificação em 2 classes (mutações que desfazem ou mantêm a interação) e a segunda em 3 classes (Se o efeito da mutação aumenta, diminui ou não afeta a afinidade do complexo proteico). As informações utilizadas foram extraídas a partir de informações obtidas pelos softwares FoldX (estrutura mutante) [Schymkowitz et al., 2005a], OPUS-PSP [Lu et al., 2008] e software GOAP [Zhou & Skolnick, 2011] referentes a parâmetros energéticos além de informações de acessibilidade ao solvente provenientes do NAccess [J. & M., 1993] dentre outros. Foram extraídas 33 atributos referentes às informações estruturais, termodinâmicas e termos extraídos de outros preditores. O coeficiente de Pearson para a regressão na base dados SKEMPI foi de 0,57 usando *Random Forest Tree* no modelo semi-supervisionado. O trabalho também fez testes com a base do 26º CAPRI (Critical Assessment of PRediction of Interactions) [Janin et al., 2003] mostrando ser o melhor preditor.

Uma outra ferramenta para a predição da variação da afinidade entre complexos é o BeAtMuSiC [Dehouck et al., 2013] utiliza redes neurais para computar a variação da energia livre dos seguintes termos da Equação 2.1.

$$\Delta\Delta G_{Bind} = \Delta\Delta G_C - (\Delta\Delta G_{P_1} + \Delta\Delta G_{P_2}) \tag{2.1}$$

No qual $\Delta\Delta G_{Bind}$ representa a variação de afinidade de complexos, considerando o complexo selvagem e o mutante, $\Delta\Delta G_C$ a variação de energia livre de enovelamento do complexo e os termos $\Delta\Delta G_{P_1}$ e $\Delta\Delta G_{P_2}$ a energia de Gibbs para o enovelamento de cada uma das partes do complexo. Assim, a correlação de Pearson para 2.007 mutações do tipo *missense* encontradas na base de dados SKEMPI teve uma correlação de Pearson de 0,40.

O mCSM proposto por [Pires et al., 2014] apresenta um modelo baseando-se na extração de informações apenas da estrutura selvagem da mutação sendo

extraídos apenas os resíduos próximos à posição em que houve a mudança de aminoácido, e, no processo de construção da assinatura é aplicado o algoritmo aCSM [Pires et al., 2013] e computada a distribuição cumulativa da matriz gerada pelo algoritmo. Juntamente com os dados da estrutura do grafo também é utilizado a diferença de atributos do farmacóforo do mutante e do selvagem. Esta estratégia atingiu uma correlação de Pearson de 0,80 para a predição da variação de afinidade proteína-proteína (sendo a mesma base utilizada em [Dehouck et al., 2013]) mas também foi utilizada para predição de termoestabilidade (Pearson de 0,82) e afinidade proteína-DNA (Pearson de 0,67). Além de ser o primeiro modelo computacional a prever efeito de mutações em afinidade proteína-DNA é o modelo com melhor correlação de Pearson para afinidade proteína-proteína existente na literatura.

Um outro trabalho que apresenta o uso de diferentes tipos de informações para esta tarefa é o BindProf [Brender & Zhang, 2015]. Nele, os autores utilizam informações de diferentes níveis da estrutura para predizer o efeito de uma mutação. A base utilizada neste trabalho foi extraída do SKEMPI, mas diferente do que foi utilizado por [Dehouck et al., 2013] e [Pires et al., 2013] neste apenas mutações que ocorrem em dímeros são utilizadas, assim, a base treinada e testada tem um conjunto de mutações é menor do que a utilizada pelos demais. O destaque do trabalho dá-se o fato do aprofundado estudo feito em relação dos diferentes tipos de atributos e como estes estão relacionados à natureza da mutação onde é comprovado que o perfil da interface (os tipos de resíduos, a posição do resíduo mutante entre outros) é crucial para o processo de predição.

O software MutaBind proposto por [Li et al., 2016] apresenta um modelo que utiliza as informações de potências energéticos utilizando o módulo ENERGY do programa CHARMM [Brooks et al., 1983], termos energéticos do NAMD [Phillips et al., 2005], elementos de estrutura secundária, área da interface calculada pela diferença da acessibilidade ao solvente, número de pontes de hidrogênio formada na interface além da hidrofobicidade do mutante e do selvagem. Com uma correlação de Pearson de 0,78 e 0,86 para um subconjunto de mutações mudando apenas um único resíduo (diferente da base proposta por [Dehouck et al., 2013]) e mutações apenas com complexos de inibidores de protease na qual cada um destes conjuntos de mutações possuem 1.925 e 862 mutações respectivamente. Uma das

limitações desta ferramenta é que somente são aceitas substituição de aminoácidos que estejam na interface.

Atualmente a principal preocupação das ferramentas é identificar o impacto da mutação baseando-se no valor da variação de energia livre de ligação entre complexos proteicos sendo que existem casos em que o trabalho foca em situações específicas encontradas na base, assim, espera-se neste trabalho contribuir para a literatura com um modelo que não apenas identifique o impacto da mutação na afinidade entre os complexos mas também apresentar uma interface capaz de representar o possível efeito da substituição de resíduos utilizando uma base de mutações mais abrangente possível.

Capítulo 3

Materiais e Métodos

A abordagem aqui apresentada é de um arcabouço para a predição da variação de estabilidade proteína-proteína, MutaGraph. A Figura 3.1 apresenta o esquema do processamento do modelo aqui proposto.

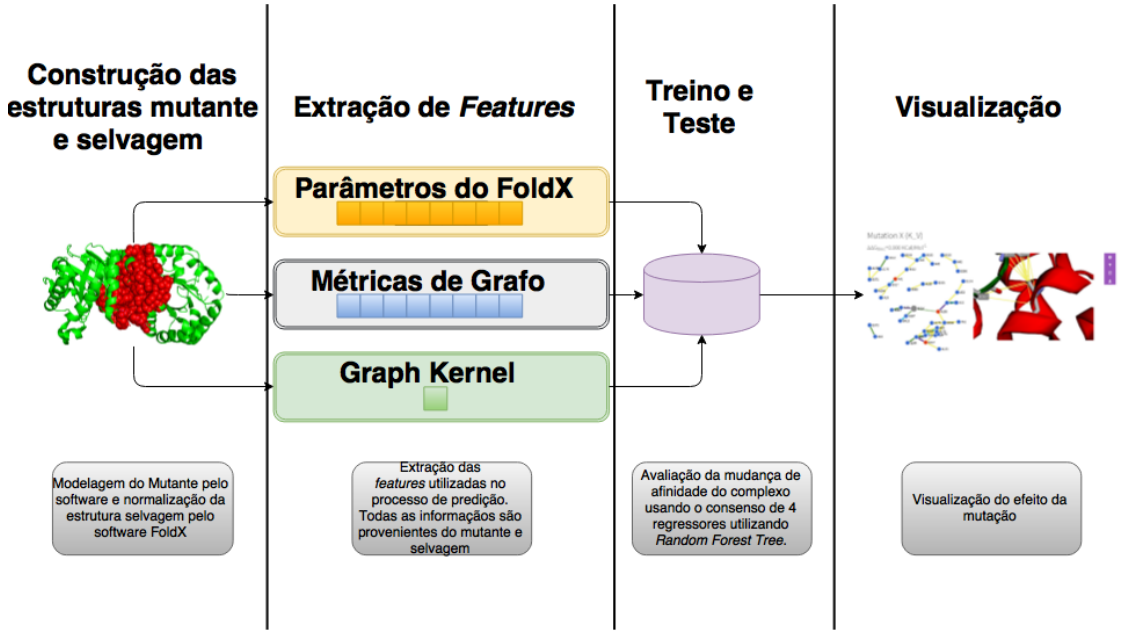


Figura 3.1: Esquema do processamento do modelo. As etapas de construção da estrutura mutante e selvagem consistem em gerar a estrutura normalizada do Selvagem (adicionar átomos faltantes) e construir a mutante pelo software FoldX. Já a etapa de extração de *features* computa todos os atributos utilizados pelos algoritmos de regressão e para visualização. A etapa de treino e teste farão a predição da mutação baseado nas informações contidas na base de dados SKEMPI, e, por fim, a visualização disponibiliza ao usuário uma interface para interagir com o resultado das etapas anteriores.

Como apresentado no esquema, podemos dividir o processamento em 3 etapas distintas descritas a seguir.

- **Pré-processamento**: Nesta etapa é feito o *download* de todas as estruturas selvagens e tratamento das mesmas tanto para a extração dos atributos quanto para a construção das estruturas mutantes utilizando modelagem por homologia pelo software FoldX [Schymkowitz et al., 2005a].

- **Processamento**: Modelagem dos dados usando rede complexa uma para as estruturas selvagens e para as estruturas mutantes bem como a extração de características tanto da rede quanto dados físico-químicos dos resíduos.

- **Construção do modelo de regressão**: Nesta etapa é feita a avaliação do modelo a ser elaborado para a predição da variação da afinidade entre proteínas usando métodos de validação cruzada (técnica apresentada na Seção 3.8) para avaliar o desempenho e por fim a construção do modelo de regressão para uso.

3.1 Conjunto de dados SKEMPI

Para este trabalho será usada a base de dados SKEMPI (Structural Kinetic and Energetic database of Mutant Protein Interactions) construída por [Moal & Fernández-Recio, 2012] com um total de 3.047 mutações extraídas da literatura com estrutura resolvida e catalogada no banco de dados PDB [Bank, 1971]. Foram extraídas um subconjunto de dados utilizadas no estudo de [Dehouck et al., 2013] em que foram utilizadas 2.007 mutações com a substituição de apenas 1 único aminoácido compreendendo um conjunto de 81 estruturas selvagens depositadas no PDB.

Alguns PDBs possuem inconsistências em sua estrutura, ou seja, resíduos faltantes ou que não são padrão além de alguns átomos de hidrogênio (considerando que nem todos os resíduos possuem esta notação). Segundo o software PDBest [Gonçalves et al., 2015] as estruturas selvagens que possuem estas inconsistências são apresentadas na Tabela 3.1.

Tabela 3.1: Relação de estruturas depositadas no PDB com incosistências.

Incosistência	**Quantidade estruturas**	**Porcentagem**
Resíduo ausentes	53	65%
Resíduo não-padrão	2	2%
Átomos ausentes	0	0%
Demais estruturas (Sem inconsistência)	27	33%

Os trabalhos [Li et al., 2016, Brender & Zhang, 2015] utilizam subconjuntos do SKEMPI no qual o conjunto de mutações utilizadas difere aos treinados e testados pelos modelos do BeAtMuSiC e mCSM de modo que a comparação entre todos os trabalhos com um único modelo de treino e teste se torna inviável, sendo assim, este trabalho irá dividir o conjunto de mutações em 3 cenários distintos condizente com cada um dos trabalhos correlatos.

- O primeiro cenário é formado pelo conjunto de mutações não-redundantes apresentadas anteriormente com 2.007 mutações, sendo esta utilizada nos trabalhos [Dehouck et al., 2013, Pires et al., 2014] sendo que este conjunto de mutações é o utilizado para a construção do modelo de predição disponível na ferramenta *web*.

- Subconjunto composto por 1.925 mutações encontradas na base de dados SKEMPI utilizado pelo software MutaBind para a construção de seu modelo (neste trabalho será chamado de SKEMPI_MutaBind para fins de diferenciação).

- A base de dados SKEMPI também possui 825 mutações relativas à estruturas com função de inibição de protease, este conjunto também é utilizado pelo MutaBind e será chamado de SKEMPIpi para fins de diferenciação.

Em relação ao conjunto de mutações utilizadas neste trabalho a Figura 3.2 apresenta a distribuição dos valores de $\Delta\Delta G$ de ligação extraídos da base de dados SKEMPI.

O maior e menor valor de energia livre de ligação encontrados na base são de 12,34 e -3,79 $KCal/mol^{-1}$ respectivamente sendo que apenas 486 mutações

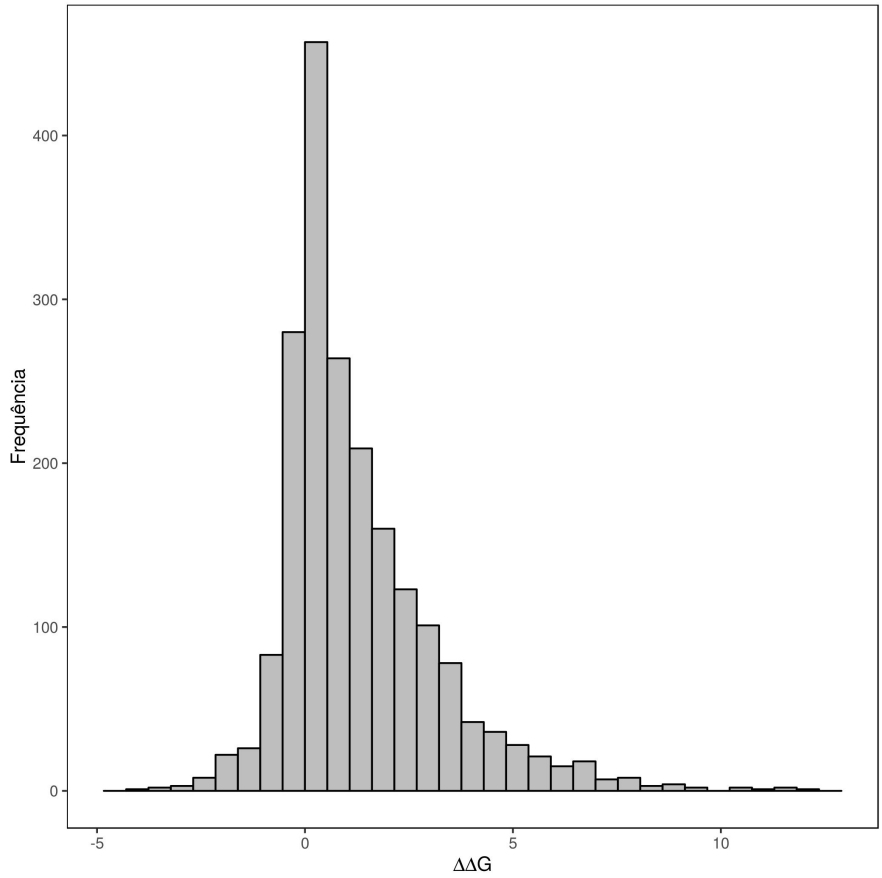


Figura 3.2: Histograma dos valores de energia livre de ligação das mutações extraídas da base de dados SKEMPI utilizadas neste trabalho.

(apenas 24% da base) possui um valor de $\Delta\Delta G$ negativo (ou seja, mutações que aumentam a afinidade do complexo) enquanto os demais registros são de perda de afinidade.

3.2 Construção das estruturas mutantes

Nem todas as mutações possuem estruturas mutantes depositadas no PDB e para a sua construção foi utilizado o software FoldX [Schymkowitz et al., 2005a] (no trabalho usou-se a versão 4) no qual cria a estrutura mutante baseada na estrutura selvagem sendo que a conformação da cadeia lateral é otimizada afim de se parecer com o mais próximo possível de um caso real, logo em seguida é feito um

refinamento da estrutura com dinâmica molecular de todos os átomos do resíduo modificado.

3.3 Variação da energia livre de ligação $\Delta\Delta G_{bind}$

[Pires et al., 2014] converteu a afinidade do complexo proteína-proteína de molar para $Kcal/mol^{-1}$ usando a seguinte fórmula de energia livre de Gibbs (ΔG).

$$\Delta G = RTln(K_D) \tag{3.1}$$

Onde R representa a constante ideal dos gases perfeitos ($R = 8.314JKmol$), T a temperatura em Kelvin e K_D a afinidade do complexo proteína-proteína em mol e o cálculo da variação da afinidade entre mutante e selvagem foi feito pela diferença entre a variação da energia livre de Gibbs de ambos como mostra a Equação 3.2.

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wildtype} \tag{3.2}$$

Em que $\Delta G_{wildtype}$ e ΔG_{mutant} representam o valor da energia livre de Gibbs da proteína em sua estrutura selvagem (aquela encontrada na natureza) e a mutante (estrutura com aminoácido modificado) respectivamente em $Kcal/mol^{-1}$. Vale ressaltar que a variação da energia livre de ligação é o valor alvo da predição do modelo aqui proposto, de forma que um valor negativo indica uma diminuição da afinidade do complexo (valor <0 $Kcal/mol^{-1}$) enquanto um valor positivo indica um aumento da afinidade no complexo proteico.

3.4 Construção da rede de contatos

A priori, para a construção da rede de contatos foi utilizado apenas contatos intercadeias. As cadeias de uma interface podem ser divididos em 2 grupos onde cada um deles deve possuir ao menos uma cadeia. Como exemplo, o SKEMPI possui as mutações do PDB 1CHO (Que é um inibidor de protease) a interface é consistida pelas cadeias E, F e G que fazem contato com a cadeia I desta estrutura. Como apresentado na Figura 3.3 apresenta esta ideia de como o conjunto de cadeias forma

a interface EFG_ I no qual os contatos entre a cadeia com as outras 3 formam a interface (contatos entre as cadeias E, F e G não são serão consideradas).

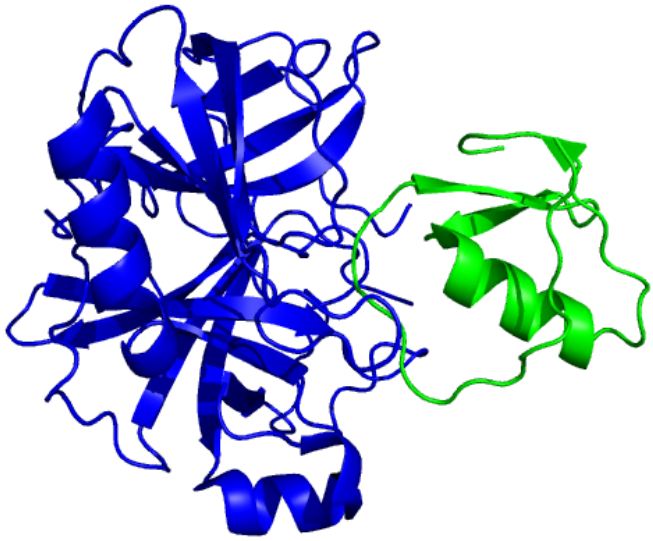


Figura 3.3: Representação da interface a ser analisada. As cadeias E, F e G estão coloridas de azul enquanto a cadeia I de cor verde. Assim, a interface considerada para análise é a feita entre a cadeia I com qualquer uma das outras 3.

Baseando-se nas premissas de tipos de contatos proposta por [Sobolev et al., 1999], a construção da rede de contatos dá-se por um algoritmo onde cada átomo da cadeia lateral é comparada com outro da cadeia vizinha a este, estando dentro de um *cutoff* especificado então o contato pode ser registrado como descrito a seguir:

1. **Pontes de hidrogênio:** Interação entre determinados átomos de oxigênio e nitrogênio mediados por um átomo de hidrogênio tendo a distância entre o aceptor e doador entre 2 à 3,2Å;

2. **Ligações hidrofóbicas:** Aminoácidos com cadeia lateral apolar (Leucina, Isoleucina, Alanina, Valina e Fenilalanina) que estejam em distâncias entre 2 e 8,6Å.

3. **Pontes salinas:** Interação devida à eletronegatividade da cadeia lateral podendo ser classificada em 2 tipos. A primeira, chamada por repulsiva

caracteriza-se quando 2 átomos de mesma carga estando numa distância entre 2 e 6Å enquanto a repulsiva (ou seja, 2 átomos de cargas diferentes) deve estar numa distância de até 8Å.

4. **Ponte dissulfeto:** Formado apenas pela interação entre 2 cisteínas tendo os seus átomos de enxofre numa distância entre 1,5 à 2.8Å do centro destes átomos.

5. **Contato aromáticos:** Os resíduos que possuem anel (Histidina, Treonina, Fenilalanina, Triptofano) devem ter os átomos do anel de um resíduo para outro numa distância entre 3 e 8Å, nesta métrica não há classificação quanto ao tipo de interação entre os anéis (ou seja, se é do tipo *edge-edge*, *face-face* ou *edge-face*).

6. **Sem classificação:** Caso o contato não se encaixe em nenhuma das características apresentadas anteriormente e a distância entre o centro destes átomos é menor que 6Å então o mesmo é considerado um contato sem classificação, pois, a proximidade entre ele o outro átomo existe, mas não se configura numa interação química.

Considerando isto, o Algoritmo 1 é executado afim de extrair estas informações tanto da estrutura mutante quanto a do selvagem.

3.5 Caracterização da rede de contatos

Como apresentado na Seção 1.4.2 os diferentes tipos de redes e afim de caracterizar e definir qual tipo de rede será analisada a Figura 3.4 apresenta a correlação entre a distância geodésica pelo coeficiente de agrupamento das redes utilizadas neste trabalho.

Como apresentado por [Barrat & Weigt, 2000] uma das características de redes *small-world* é sua correlação ser intermediária às redes regulares e aleatórias. Sendo assim, a sua alta distância geodésica em relação ao seu baixo coeficiente de agrupamento faz com que fenômenos de redes livres de escala (*hubs* que, se removidos, o grafo se torna desconexo), redes regulares (existe a mesma quantidade

Algoritmo 1: Pseudocódigo do algoritmo da construção do grafo de contatos entre interfaces.

```
Entrada: Estrutura PDB e conjunto de cadeias que forma a interface
         (cs_1 e cs_2).
G = Grafo não-direcionado vazio;
para cada cadeia α existente em cs_1 faça
    para cada cadeia β existente em cs_2 faça
        para cada átomo da cadeia lateral (w) dos resíduos de α faça
            para cada átomo da cadeia lateral (v) dos resíduos de α faça
                se Distância entre w e v for menor que cutoff então
                    Registrar os átomos w e v como vértices em G;
                    Adicionar a aresta entre w e v no grafo G rotulando-a
                     partir do tipo de contato;
retorna G;
```

de arestas para todos os vértices) e fenômenos de redes aleatórias não são esperados neste modelo.

3.6 Extração de *features*

3.6.1 Métricas de topologia de rede

Com o objetivo de saber a importância de cada resíduo dentro da rede algumas métricas foram baseadas no trabalho [Li et al., 2012] onde diferentes tipos de atributos de grafos são extraídos para classificação, ainda, o próprio trabalho utiliza como teste para avaliar a qualidade destes atributos uma base de compostos químicos, interação entre complexos proteicos e grafo de interação entre diferentes tipos de tecidos.

O conjunto de atributos extraídos das redes formadas pelas interfaces da estrutura mutante e selvagem são descritas a seguir.

1. **Grau médio**: O grau de um nó é definido pela quantidade de arestas que o ligam até os seus vizinhos enquanto a soma do grau de todos os vértices divi-

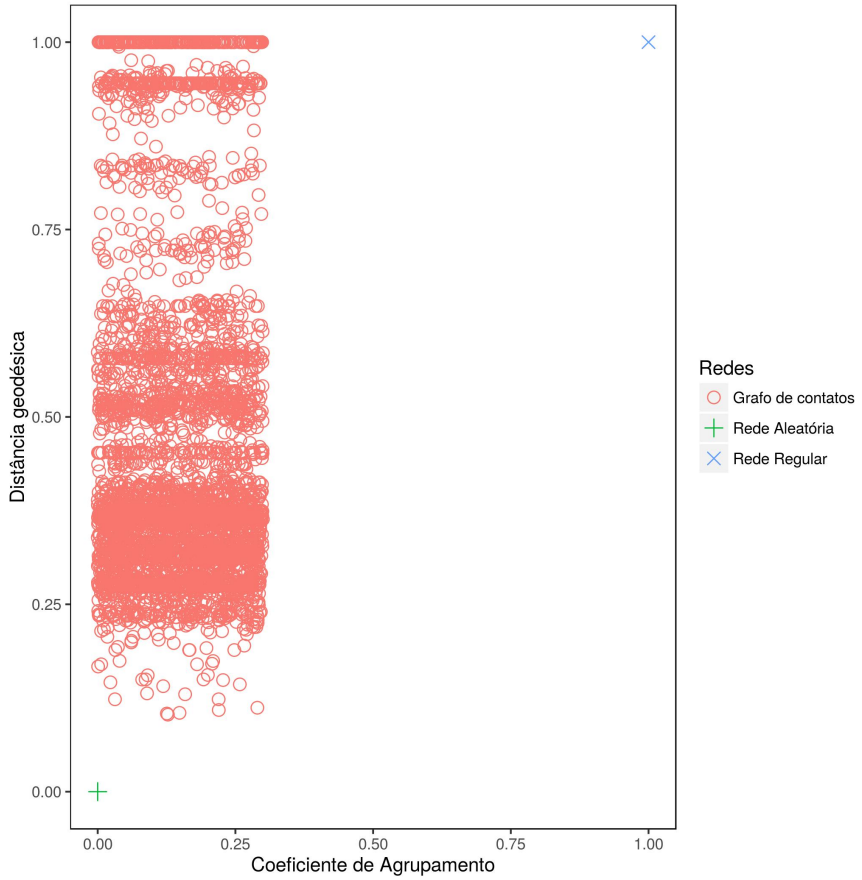


Figura 3.4: Representação da distância geodésica e coeficiente de agrupamento (ambos normalizados) das 2.088 estruturas utilizadas neste trabalho (2.007 estruturas mutantes mais 81 estruturas) juntamente com a relação de exemplo de uma rede Aleatória e outra Regular para fins de referência.

dido pela quantidade de nós do grafo define a sua média, como apresentado na Equação 3.3 onde $d(u_i)$ é o grau de um nó qualquer.

$$\hat{d}(G) = \sum_{i=1}^{n} \frac{d(u_i)}{n} \tag{3.3}$$

2. **Média do coeficiente de agrupamento**: Dado um nó u qualquer consideramos o seu coeficiente de agrupamento $c(u)$ é a probabilidade de um determinado nó ter o seus vizinhos como um grafo clique, em outras palavras, esta métrica é responsável por avaliar a densidade de ligação dos nós vizinhos ao

avaliado. A Equação 3.4 apresenta o cálculo do coeficiente de agrupamento para um único nó tendo $\lambda(u_i)$ é o número de triângulos formado por seus vizinhos enquanto $\tau(u_i)$ é o número de possíveis grafos cliques que podem ser formados, e, na Equação 3.5 temos o coeficiente médio de agrupamento ($C(G)$) sendo esta apenas a média aritmética do coeficiente de agrupamento de todos os vértices.

$$c(u) = \frac{\lambda(u_i)}{\tau(u_i)} \tag{3.4}$$

$$C(G) = \frac{\sum_{i=1}^{n} c(u_i)}{n} \tag{3.5}$$

3. **Média da excentricidade**: A excentricidade de um vértice dá-se pelo maior menor caminho do vértice em questão em relação aos demais. As Equações 3.6 e 3.7 descrevem o cálculo utilizado para a excentricidade (considere $d(u, v)$ como o menor caminho entre os vértices u e v).

$$e(u) = \sum_{i=1}^{n} max\{d(u, v) : v \in V\} \tag{3.6}$$

$$\overline{e}(G) = \frac{1}{n} \sum_{i=1}^{n} e(v_i)|v_i \exists G \tag{3.7}$$

4. **Maior excentricidade**: Também pode ser chamado de maior diâmetro efetivo. Neste caso, buscamos aqui o maior menor caminho de toda o grafo.

5. **Menor excentricidade**: O menor caminho possível dentro de um grafo também pode utilizar chamado de raio efetivo de um grafo.

6. ***Closeness***: Métrica que apresenta o quão próximo um vértice está dos demais na rede. Esta métrica é importante para identificar a posição de um vértice na rede, quanto menor o closeness, mais afastado do centro o vértice estará. A Equação 3.8 apresenta a fórmula da métrica.

$$C(v) = \frac{1 - |V|}{\sum_{s \ \epsilon \ V} \sigma(s, v)} \tag{3.8}$$

Onde $|V|$ é a quantidade de vértices do grafo e $\sigma(s, v)$ o menor caminho do vértice s à v [Freeman, 1979].

7. **Porcentagem de vértices de ponta**: Nesta métrica apenas contabilizamos o número de vértices com grau 1 em razão do número total de vértices.

8. **Raio espectral e o segundo maior autovalor**: Dado a matriz de adjacência de um grafo consideramos então o raio espectral como o maior valor do autovalor da matriz de adjacências, e, junto a isto utilizamos o segundo maior autovalor computado considerando que $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$.

9. **Energia do grafo**: Basicamente é a soma do quadrado de todos os autovalores da matriz de adjacência como apresentado na Equação 3.9.

$$E(G) = \sum_{i=1}^{n} \lambda_i^2 \tag{3.9}$$

10. **Quantidade de autovalores**: É a contagem da quantidade de distintos valores dentro do autovalor. Apesar do autovalor possuir n valores (onde n é a quantidade de vértices do grafo) estes valores não necessariamente são distintos entre si.

11. **Entropia de rótulo**: Baseando-se no modelo de entropia de Shannon afim de mensurar a incerteza dos rótulos do grafo, em linhas gerais, podemos considerar esta métrica como "quanto menor a possibilidade de acertos aleatórios num experimento, menor é a capacidade em se observar a ocorrência deste evento" [Lin, 1991, Coifman & Wickerhauser, 1992]. Considerando os diferentes tipos de rótulos que um vértice u qualquer possa assumir de um conjunto l podemos calcular sua entropia como mostra a Equação 3.10.

$$H(G) = -\sum_{i=1}^{n} p(l_i) log \ p(l_i) \tag{3.10}$$

12. **Impureza da vizinhança**: Definimos a impureza de um vértice u qualquer como sendo a quantidade de vizinhos que possuem um rótulo diferente dele mesmo. Assim, a Equação 3.11 apresenta esta relação tendo como $L(u)$ e $L(v)$ são os rótulos do vértice em questão e de um do seus vizinhos respectivamente e $N(u)$ é o conjunto de vizinhos do vértice u.

$$VI(u) = |L(v) : v \in N(u), L(u) \neq L(v)| \tag{3.11}$$

13. **Impureza de aresta**: Uma aresta é considera impura se e somente se os rótulos dos vértices desta aresta são diferentes entre si, sendo assim, a impureza de arestas de um grafo G qualquer dá-se pela Equação 3.12 onde m é a quantidade de arestas do grafo.

$$EI(G) = \frac{|(u,v) \in E : L(u) \neq L(v)}{m} \tag{3.12}$$

14. ***Betweenness***: É a medida de centralidade de um determinado vértice no grafo, basicamente, é a contagem de todos os caminhos mais curtos de todos os vértices contra todos os demais que passam pelo vértice dividido pela quantidade total de caminhos mais curtos. A Equação 3.13 apresenta esta relação. Ainda, vale ressaltar que esta métrica não se encontra no trabalho de [Li et al., 2012] mas é um atributo importante no modelo.

$$B(v) = \sum_{s,t \ \epsilon \ V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{3.13}$$

Além destes atributos, ainda temos a quantidade de vértices e arestas do grafo como atributos. [Li et al., 2012] ainda apresenta outras métricas para avaliação do modelo como por exemplo o maior conjunto de vértices conectados no grafo e a porcentagem de vértices isolados (ou seja, vértices de grau 0) mas estes atributos não serão utilizados pois estas características não são encontradas nos grafos extraídos das interfaces.

3.6.2 Graph kernel

Comparar e mensurar diferença entre 2 grafos distintos é um problema computacionalmente caro. Identificar isomorfismo em subgrafos é um problema de natureza $\mathbb{NP}$-completo, e, consequentemente seu uso se torna restrito, especialmente para análises em larga escala. Levando em conta estes fatores, o *Graph kernel* pode se tornar a melhor solução para este tipo de problema.

Formalmente, podemos dizer que um *Graph Kernel* K deve ser capaz de comparar dois grafos distintos e obedecer as seguintes propriedades [Vishwanathan et al., 2010]:

- Deve ser simétrico ($k(G, G') = k(G', G)$)
- O resultado da função deve ser um valor inteiro positivo.

O principal desafio desta abordagem está em extrair de forma mais precisa possível a semântica da diferença entre 2 grafos quaisquer.

[Samatova et al., 2013] apresenta um algoritmo de *kernel* para grafos, o objetivo desta métrica é a definição de um valor adimensional com o objetivo de caracterizar a semelhança topológica entre dois grafos distintos tendo a sua aplicação em trabalhos que manipulam redes complexas [Ranshous et al., 2015, Rao & Rao, 2014] quanto em problemas envolvendo Bioinformática [Rujirapipat et al., 2017].

O primeiro passo para a construção desta assinatura é o produto notável de 2 grafos, que, no contexto deste trabalho são os contatos feitos pela posição em que ocorrerá a mutação. O primeiro passo é a construção do produto direto (também conhecido como produto de kronecker) da matriz de adjacência resultante é convertido num grafo. A Equação 3.14 apresenta o método algébrico do produto direto enquanto a Figura 3.5 representa o grafo gerado.

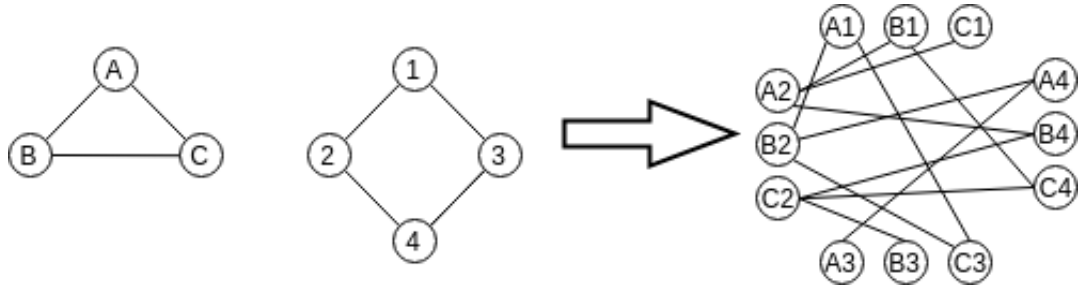


Figura 3.5: Representação do processo de produto direto entre 2 grafos. Basicamente é a aplicação do produto direto em ambas as matrizes de adjacência.

$$A_{n\times m} \times B_{r\times s} = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nm}B \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{11} & \cdots & a_{1m}b_{1s} \\ a_{11}b_{21} & a_{11}b_{22} & \cdots & a_{1m}b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}b_{r1} & a_{n1}b_{rs} & \cdots & a_{nm}b_{rs} \end{bmatrix} \tag{3.14}$$

Com o grafo produto construído, o Algoritmo 2 apresenta os passos da construção do valor numérico utilizado como atributo.

Algoritmo 2: Pseudocódigo para o computo do valor do kernel.

Entrada: Grafo produto G'
$d_i \leftarrow$ Maior grau de entrada de G';
$d_o \leftarrow$ Maior grau de saída de G';
$\gamma \leftarrow \frac{1}{min(d_i,d_o)}$;
retorna $\sum_{i,j}(I - \frac{A_ij}{\gamma})^{-1}$;

Computacionalmente, o principal problema deste algoritmo está no espaço de memória alocado. Devido o produto direto entre matrizes ser a multiplicação de cada elemento da matriz A pela matriz B isto torna seu uso restrito [Fernandes et al., 2013] e neste trabalho não foi possível avaliar um grafo maior do que aqui utilizado devido à este fato. Ainda, de todos os atributos até então extraídos este é o único em que o seu resultado utiliza tanto a estrutura mutante quanto a selvagem simultaneamente.

3.6.3 Atributos auxiliares

Além dos atributos de grafo ainda temos o conjunto de informações extraídos do FoldX [Van Durme et al., 2011] no momento em que a estrutura mutante é gerada (estes valores são computados pelo FoldX tanto para o mutante quanto para o selvagem) e os atributos com a sua descrição são apresentadas abaixo a partir da explicação feita por [Schymkowitz et al., 2005b]:

1. Solvatação de hidrofóbicos e polares: É computada a partir da quantidade de átomos enterrados juntamente com a contribuição do grupo de átomos hidrofóbicos ou polares da estrutura. Estes parâmetros de solvatação foram derivados de experimentos em que os aminoácidos são transferidos da água para um solvente orgânico, assim, assumindo que a simulação e transição que pode ocorrer em um aminoácido durante o processo de enovelamento num meio aquoso.

2. Termos de Van der Waals: Esta métrica refere-se às energias de transferência da água para o vapor (processo de dessolvatação).

3. Pontes de hidrogênio: As ligações de hidrogênio são calculadas a partir de distância geométrica simples juntamente com a sua energia.

4. Energia eletrostática: É calculada a partir de uma implementação simples da lei de Coulomb. A constante dielétrica é escalonada à partir da ligação em questão.

5. Taxa de associação do complexo: Utiliza a equação empírica de Schreiber [Selzer et al., 2000][1] que estima a taxa de associação do complexo.

6. Custo da entropia para a conformação da cadeia lateral e principal: A energia dissipada para que tanto a cadeia lateral quanto a cadeia principal entre em sua conformação e este é calculado à partir dos parâmetros apresentados por [Abagyan & Totrov, 1994][2].

[1] O método proposto basicamente aumenta a atração eletrostática entre as proteínas incorporando resíduos carregados na proximidade da interface em questão.

[2] O computo destes parâmetros é feito à partir do modelo de monte carlo com o objetivo de minimizar o espaçamento entre os ângulos ϕ e ψ da cadeia principal.

7. Sobreposição estética: Medida que calcula a sobreposição estética dos átomos da cadeia principal.

3.6.3.1 Área de contato da interface

Um outro atributo também computado é a acessibilidade ao solvente. Este representa área da estrutura exposta ao solvente mensurada em Angströms^2 sendo a soma da acessibilidade ao solvente de cada átomo resulta na acessibilidade ao solvente do resíduo.

Neste trabalho a acessibilidade ao solvente foi computada pelo software DSSP [Touw et al., 2015]. Para calcular a área de contato da interface é utilizado a Equação 3.15.

$$Asa_I = \sum_{i=1}^{n} Asa_{C_i} - Asa_T \tag{3.15}$$

Onde Asa_I é a área de contato da interface, Asa_{C_i} é a área acessível do complexo isolado enquanto Asa_T é a acessibilidade ao solvente do complexo (ou seja, considerando todas as cadeias).

Para apresentação da Acessibilidade ao solvente dos resíduos na ferramenta online utilizam a acessibilidade ao solvente relativa. Como o DSSP não nos fornece este valor foi utilizado como base o computo feito pelo NACCESS [Hubbard & Thornton, 1993] no qual utiliza os valores propostos por [Miller et al., 1987].

3.6.4 Assinatura da mutação

Com todas estas informações extraídas da mutação, cada um dos atributos até o momento presente estão relacionados na Tabela 3.2 juntamente com a sua origem (se veio do mutante ou selvagem).

Como cada um dos algoritmos apresentados possuem diferentes formas de construírem o seu modelo de decisão consequentemente um mesmo conjunto de dados para cada um deles pode ter diferentes resultados. Sendo assim, o primeiro passo foi a escolha do conjunto de atributos para cada um destes algoritmos. Foi

Tabela 3.2: Relação dos atributos extraídos das estruturas.

Grupo	**Estrutura Selvagem**	**Estrutura Mutante**	**Total**
Estrutura do grafo	18	18	36
Classificação dos contatos	7	7	14
Atributos auxiliares	22	22	44
Graph Kernel			1
Total			95

utilizado o algoritmo de eliminação de atributos por recursividade[3] para que cada um dos algoritmos utilizados possua o seu próprio conjunto de dados afim de melhorar o resultado de cada um deles [Deng & Runger, 2012, Gantz & Reinsel, 2011].

O principal objetivo deste algoritmo é criar um menor subconjunto de atributos possíveis de tal forma que maximize a qualidade do resultado obtido. O Algoritmo 3 apresenta os passos para o filtro feito dos atributos utilizado por cada um dos algoritmos de aprendizado supervisionado para regressão apresentados na Seção 3.11.

Algoritmo 3: Pseudocódigo do algoritmo de eliminação de atributos por recursividade.

Entrada: S como o conjunto de atributos e Θ como o estimador
enquanto *Existir atributos que possam ser removidos de* S **faça**
    Utilize o estimador Θ em S;
    Ordene S em função da relevância no estimador Θ;
    Remova de S o atributo com menor relevância;

Temos então Θ como o algoritmo de regressão utilizado nos atributos. Desta forma, como cada algoritmo de regressão utiliza o conjunto de dados de forma distinta notou-se então que o conjunto final de atributos utilizado em cada um deles foi distinto. O Anexo B apresenta a relação de todos os atributos em função do algoritmo que o utilizou.

[3]Do inglês: *Recursive Feature Elimination*

3.7 Redução de dimensionalidade

Com o objetivo de reduzir o ruído existente nos dados e a dimensionalidade foi utilizado o SVD (*Singular Value Decomposition*) que é uma técnica de álgebra linear para fatoração de matrizes e descoberta de características latentes [Eldén, 2006, Eldén, 2007].

O SVD busca correlação nos dados e os agrupa formando combinações lineares e reduz a dimensionalidade dos dados. Dada uma matriz $X_{n_x m}$ o SVD irá fatorá-la como apresentado na Equação 3.16.

$$X_{n\times m} = U_{n\times n}\Sigma_{n\times m}V^T_{m\times m} \tag{3.16}$$

Onde $U_{n\times n}$ e $T_{m\times m}$ são as matrizes de valores singulares da esquerda e direita enquanto $\Sigma_{n\times m}$ é matriz diagonal com os valores singulares de $X_{n\times m}$ ordenados em ordem decrescente de significância.

Ao considerarmos um subconjunto de valores singulares k onde $k < p$ e que p representa o posto original da matriz X então temos.

$$X \approx X_k = U_k\Sigma_k V^T_k \tag{3.17}$$

Como apresentado na Figura 3.6 a decomposição em valores singulares de um conjunto de dados têm por objetivo retirar as características específicas de cada um deles afim de que apenas as informações mais relevantes de cada instância se torna detectável.

A versão do software R foi utilizada neste trabalho para o processamento das matrizes [Becker et al., 1988] e o método para a redução de posto é a mesma proposta por [Eldén, 2006]. O Algoritmo 4 apresenta os passos feitos para a redução de posto dos conjuntos de treino e teste para avaliação do modelo.

3.8 Avaliação do modelo

Com o objetivo de evitar um modelo tendencioso ou enviesado foi utilizado o método de validação cruzada de 10 ou K dobras. Segundo [Zaki & Wagner Meira, 2014] este método divide o conjunto de dados em k (no

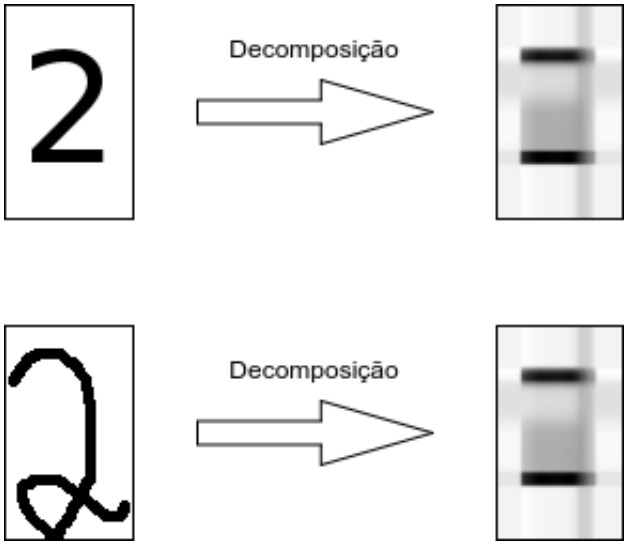


Figura 3.6: Decomposição de 2 imagens distintas com o algarismo 2. No final do processo de decomposição de ambas as imagens estas ficaram mais semelhantes entre si do que as originais.

Algoritmo 4: Pseudocódigo da redução de posto pela técnica de SVD proposta por [Eldén, 2006].

Entrada: Matriz de treino (Λ), a matriz de teste (Θ) e o posto de interesse k

$U, \Sigma, V^T = \text{SVD}(\Lambda^T)$;

$\Lambda' = (\Sigma_k V_k^T)^T$;

$\Theta' = \Theta U_k$;

retorna Λ' *e* Θ';

caso 10) partes iguais chamadas de dobras. Cada i-ésima dobra é utilizada como teste enquanto as demais são utilizadas para treino em um determinado cenário. Geralmente os valores atribuídos à k são 5 ou 10. Neste trabalho foram utilizadas 10 dobras e em cada cenário uma determinada dobra é utilizada como teste enquanto as demais são utilizadas como treino no processo de regressão.

Para que as dobras da validação cruzada não construa um cenário desbalanceado foi utilizado a técnica de dobras estratificadas onde o objetivo é manter a distribuição dos dados semelhantes tanto no conjunto de treino quanto no de teste.

3.8.1 Validação cruzada de baixa redundância

Além da validação cruzada comum também foi usado o modelo de baixa redundância proposto por [Pires et al., 2014]. Com o objetivo de criar um processo de validação cruzada que evite vieses na regressão construída todas as mutações que ocorrem na mesma posição de uma proteína são agrupadas na mesma dobra.

3.9 Métricas de avaliação

3.9.1 Coeficiente de correlação de Pearson

Utilizada para mensurar o nível de associação entre duas variáveis. O resultado desta métrica (valor de ρ) pode ser interpretado da seguinte forma:

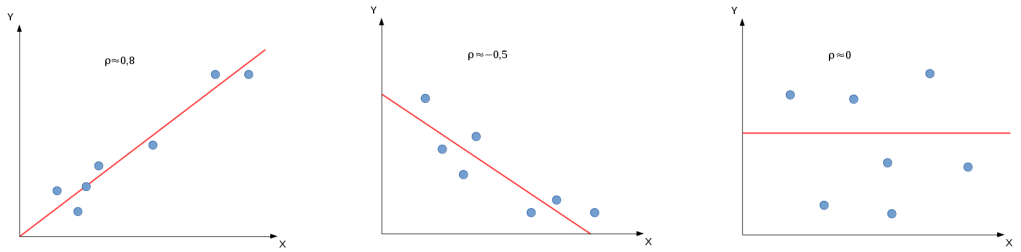
1. Se $\rho = -1$ significa uma correlação inversamente perfeita;
2. Se $\rho = 0$ significa que não existe correlação linear entre as variáveis;
3. Se $\rho = 1$ significa uma correlação perfeita.

A Equação 3.18 apresenta a fórmula utilizada para esta métrica. O objetivo do uso neste trabalho é averiguar a qualidade do algoritmo de regressão podendo analisar pelo modelo de validação cruzada a correlação entre o $\Delta\Delta G$ de uma mutação com o encontrado no processo de regressão. A Figura 3.7 apresenta graficamente a correlação linear nos 3 casos apresentados.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.18}$$

3.10 Teste estatístico para significância de diferenças de correlação

Com o objetivo de mensurar se determinadas diferenças nos resultados referentes à correlação de Pearson encontradas neste trabalho são significantes foi



(a) Coeficiente de pearson para $\rho > 0$.

(b) Coeficiente de pearson para $\rho < 0$.

(c) oeficiente de pearson para $\rho = 0$.

Figura 3.7: Representação gráfica do coeficiente de Pearson para três casos distintos. Em (a) temos uma correlação direta, em (b) temos uma correlação indireta e em (c) quando não há correlação linear.

utilizado o teste estatístico proposto por [Steiger, 1980] no qual é feito um teste de igualdade entre duas correlações de Pearson a partir da mesma amostra tendo como resultado um *z-score* comparado numa distribuição bilateral [Hoerger, 2013, Lee & Preacher, 2013].

3.11 Processo de regressão

Um dos objetivos do trabalho é predizer em função do conjunto de atributos extraídos com a variação da energia livre de ligação ($\Delta\Delta G_{bind}$) levando em conta os atributos extraídos. Para isso, foi feito o consenso entre diferentes algoritmos de forma que o resultado predito por estes é a entrada de outro que fará o consenso de forma a melhorar o resultado [Alpaydin, 2007]. A Figura 3.8 apresenta o esquema do uso desta técnica.

A partir desta ideia, foram utilizados 4 algoritmos diferentes no processo de regressão num primeiro momento, e, afim de tornar o modelo mais,preciso possível uma nova etapa de regressão é feita onde o conjunto de atributos deste nada mais é que o valor estipulado pelos 4 primeiros. As Seções a seguir apresentará a estratégia que cada algoritmo utiliza. Uma vez que cada um deles tem diferentes formas de abordar o mesmo problema.

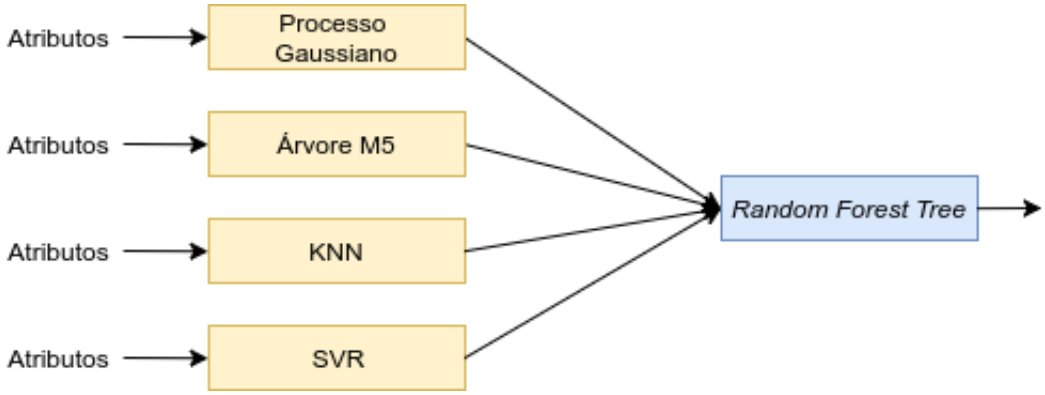


Figura 3.8: Representação do modelo de regressão por consenso para quatro diferentes algoritmos (Processo Gaussiano, Árvores M5, KNN e SVR) nos quais o resultado destes formarão o conjunto de atributos da árvore de regressão.

3.11.1 Processo Gaussiano

Diferente de uma regressão linear onde devemos encontrar os melhores coeficientes possíveis para as variáveis e um valor para o termo independente de uma função linear para o nosso conjunto de dados, o Processo Gaussiano (GP) trabalha numa abordagem não paramétrica[4] de forma que o mesmo começa com uma distribuição inicial e vai ajustando-a partir dos pontos observados no modelo de teste [Murphy, 2012].

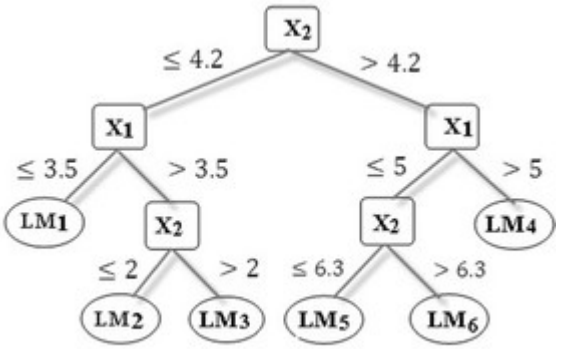
Temos então o processo gaussiano como um conjunto de dados n onde cada atributo $(x_1, x_2, \cdots, x_n)$ é uma distribuição conjunto de variáveis aleatórias $(f(x_1), f(x_2), \cdots, f(x_n))$ tendo a média $\mu(x)$ e a sua covariância baseada numa função de kernel como ilustrado na Equação 3.19 onde $f(x)$ é a variável aleatória de resposta que posteriormente é convertida para o espaço de atributos.

$$f(x) \sim GP(\mu(x), K(x, x')) \tag{3.19}$$

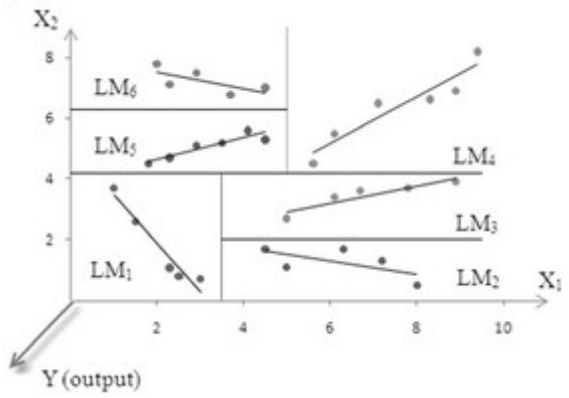
3.11.2 Árvore de Regras M5

Também conhecida como M5P é uma estrutura de decisão para tarefas de regressão para a predição da variável de interesse. Similar à ideia do uso de distintos atri-

[4]Ou seja, independe de características intrínsecas aos dados.



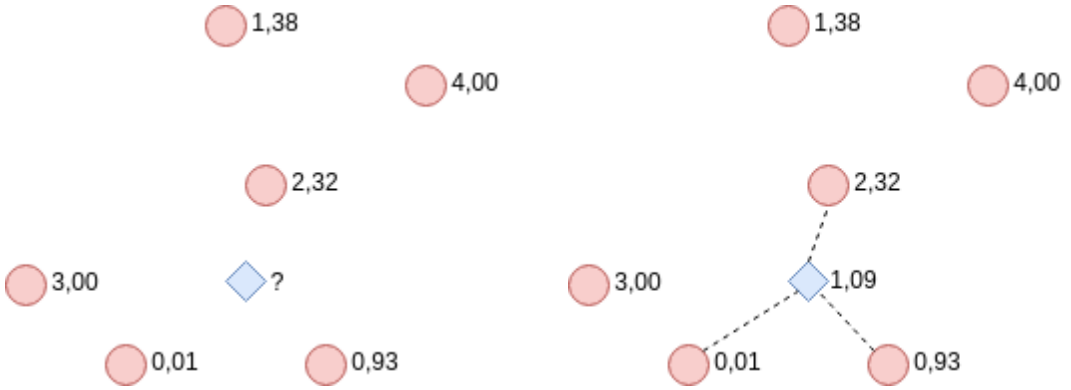
(a) Árvore de regressão



(b) Espaço bidimensional separado pelas regras da árvore

Figura 3.9: Árvore de regras M5 (a) para um conjunto de dados num espaço bidimensional tendo uma terceira dimensão como a variável de interesse (b) [Rahimikhoob et al., 2013].

butos para a construção do modelo da árvore para decidir em qual classe pertence numa árvore de decisão, cada folha neste caso possui uma função linear tornando assim o modelo de regressão linear multivariada. Segundo [Quinlan et al., 1992] esta abordagem torna possível a precisão em modelos de alta dimensionalidade e ainda a árvore gerada tende a ser menor do que algoritmos semelhantes. A Figura 3.9 apresenta a árvore de regras M5 construída para um conjunto de dados.



(a) Instância de interesse no conjunto de teste.

(b) Valor computado da média das 3 instâncias mais próximas da instância de interesse.

Figura 3.10: Execução do algoritmo KNN. O losango representa a instância de interesse enquanto os círculos vermelhos as instâncias do conjunto de Treino. Em 3.10b temos a instância com a variável de interesse é predita pela média dos 3 instâncias mais próximas.

3.11.3 k-Nearest Neighbors

Este é um algoritmo simples de aprendizagem onde basicamente dado uma entrada que deseja-se mensurar sua variável de interesse a partir da similaridade da instância em questão com o seu conjunto de treinamento.

No processo de regressão escolhe-se as K instâncias mais próxima da instância sendo analisada e para o processo de regressão faz-se a média entre as variáveis de interesse como ilustrado na Figura 4.1.

3.11.4 Vetores de suporte de Regressão (SVR[5])

No início da década de 60, o algoritmo de suporte de vetores foi criado com o objetivo de construir hiperplanos separadores para problemas de reconhecimento de padrões [Vapnik & Lerner, 1963, Gaines & Andreae, 1966]. Apesar disso, apenas na década de 90 foi generalizado para construir funções separadoras não-lineares [Boser et al., 1992] e estimar valores reais (Regressão) [Cortes & Vapnik, 1995].

[5]Do inglês *Support Vector Regression*

A máquina de suporte vetorial utiliza um processo de aprendizado baseando-se num conjunto de treino usando teoria estatística de aprendizagem usando vetores de espaço de entrada não mapeados não linearmente para um espaço característico de alta dimensionalidade através de um mapeamento escolhido *a priori*.

O principal objetivo do Suporte de vetores para regressão usa uma função de perda para os problemas de regressão, ou seja, erros no modelo serão aceito mas não maior que um valor estipulado (ϵ) criando vetores com o menor erro possível e ao mesmo tempo o mais flexível possível. Em alguns casos é comum encontrar instâncias que fiquem fora da margem de erro, ainda, o parâmetro C do modelo é usado para compensar o a habilidade de generalização do modelo com a acurácia do conjunto de treino [Crone et al., 2006]. A Figura 3.11 apresenta um esquema de construção do modelo de regressão construído pelo SVR.

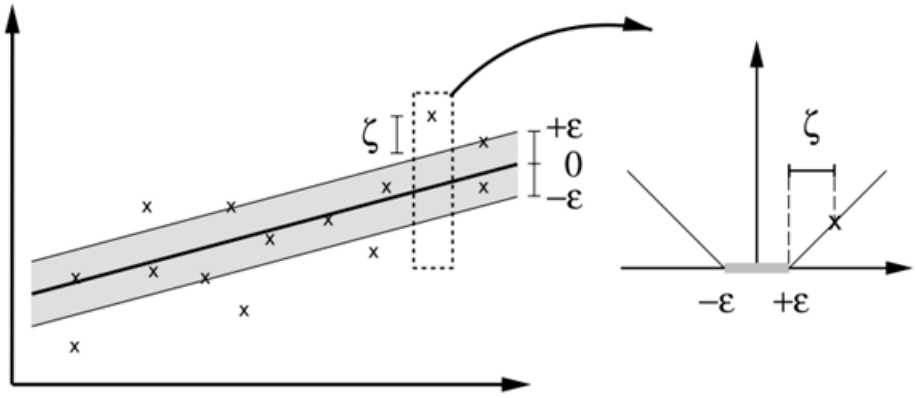


Figura 3.11: Construção do modelo de regressão criado pelo SVR mostrando o uso das variáveis de erro (ϵ) e de custo (ζ) para a construção da margem de perda [Schölkopf et al., 1998].

Assim, o objetivo do uso das máquinas de suporte é para a construção do modelo de regressão tendo como variáveis de entrada as métricas de redes complexas juntamente com os dados biológicos afim de predizer a variação da energia livre de ligação ($\Delta\Delta$G), uma vez que a correlação entre as variáveis de entrada com a de interesse não possui correlação linear espera-se que o SVR construa um modelo que encontre uma correlação relevante para estas variáveis.

3.11.5 *Random Forest Tree*

O Algoritmo de *Random Forest Tree* é utilizado tanto para classificação quanto para regressão e este é eficiente pois sua premissa é a junção de diferentes modelos de baixa qualidade para construir um modelo mais preciso. Outras características relevantes deste algoritmo está na sua capacidade de trabalhar com valores faltantes, redução de dimensões e conjunto de valores desproporcionais [Chan & Paelinckx, 2008, Schroff et al., 2008, Svetnik et al., 2003].

A principal ideia do algoritmo é a geração de diferentes árvores que darão um resultado para a regressão ou classificação dependendo do problema em questão. Basicamente é feito um *bootstraping* com todas as instâncias de treino, ou seja, o algoritmo irá tentar construir várias árvores de decisão com diferentes amostras e variáveis na raiz de cada uma delas. O resultado final dá-se por votação me problemas de classificação e a média de cada árvore num processo de regressão. A Figura 3.12 apesenta o esquema do processo de construção do modelo de previsão do algoritmo *Random Forest Tree*.

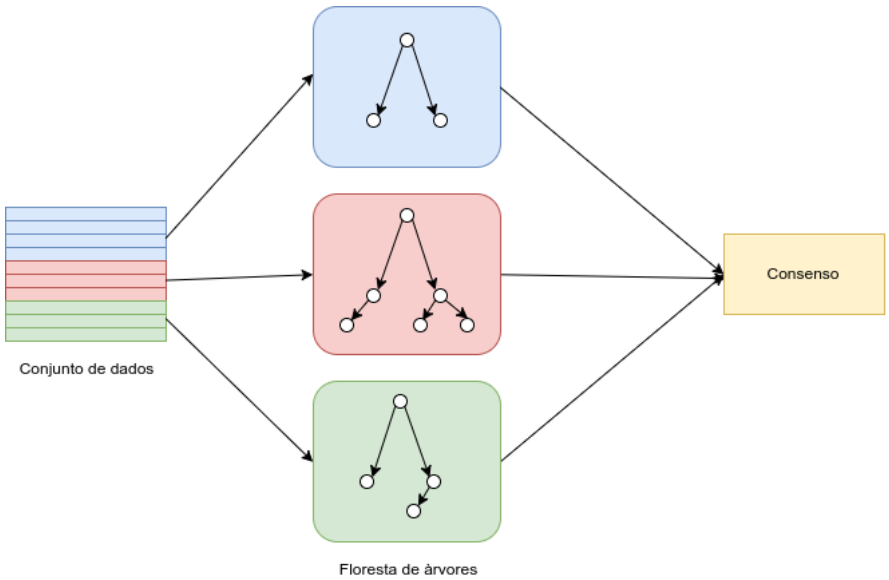


Figura 3.12: Esquema de construção de floresta de *Random Forest Tree*. O conjunto de dados é separado em diferentes conjuntos de amostras afim de gerar um conjunto de árvore. Por fim o consenso entre o conjunto de cada uma delas dará o resultado final do modelo.

Capítulo 4

Resultados e discussões

As Seções a seguir irão apresentar os resultados obtidos com o modelo apresentado anteriormente. Num primeiro momento será apresentado a construção do modelo de regressão mostrando os resultados no método de escolha do K para o algoritmo de KNN passando pela seleção de atributos para cada um dos algoritmos utilizados no sistema e por fim, a estratégia de redução de dimensionalidade utilizando SVD para melhorar os resultados até a construção do modelo de regressão final.

Nesse capítulo, apresentamos uma comparação desse trabalho com os demais da literatura que tratam do problema de predição do efeito de mutações em interações proteína-proteína, e, por fim um estudo de caso utilizando a ferramenta online para visualização do efeito da mutação.

4.1 Construção do modelo de Regressão

4.1.1 Seleção do K para o algoritmo KNN

Diferente dos demais algoritmos utilizados neste trabalho, o KNN possui uma peculiaridade em relação aos demais. A quantidade de vizinhos mais próximos altera consideravelmente a qualidade do resultado mas ao mesmo tempo existe uma relação direta entre eles [Deng et al., 2016]. Em outras palavras, é possível avaliar o ganho na qualidade do resultado do algoritmo KNN ao aumentar a quantidade de vizinhos selecionados para computar o resultado. Este fenômeno não pode

ser observado por outros algoritmos já que para alguns deles é necessário outras técnicas para este fim, como por exemplo, o uso de algoritmos genéticos para encontrar a melhor combinação dos parâmetros do SVR [Wu et al., 2009].

Como não é possível definir *a priori* qual a melhor quantidade de vizinhos a serem utilizados para o processo de regressão com o algoritmo KNN. O primeiro passo foi variar o valor de K entre 1 e 25 afim de escolher o melhor valor como apresentado na Figura 4.1

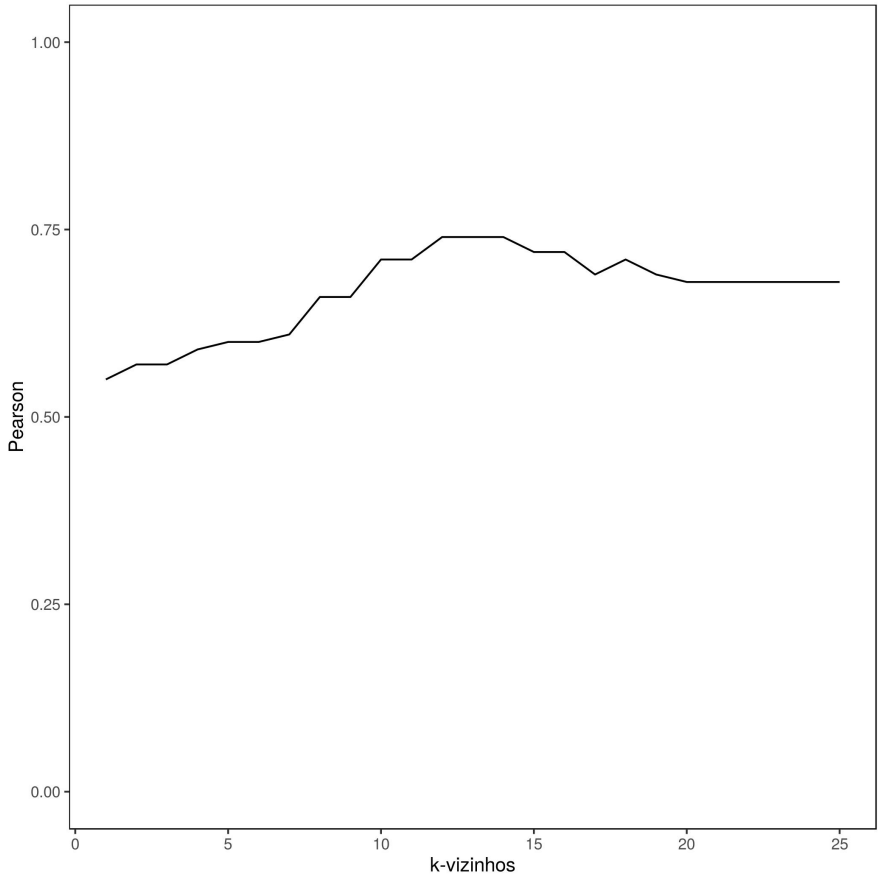


Figura 4.1: Correlação de Pearson fazendo a regressão com o algoritmo KNN, o valor de K variando entre 1 e 25.

Como observado no gráfico o melhor valor para K foi 12 primeiros elementos (assumindo o uso de todos os atributos) uma vez que a partir deste ponto ao aumentar o número de K a qualidade do processo de regressão perdendo a qualidade.

Para o computo da variação da energia livre de ligação de uma mutação é feito a média dos 12 primeiros elementos mais próximos da instância comparada.

4.1.2 Caracterização dos atributos utilizados

Como apresentado na Seção 3.6 4 grupos distintos de atributos foram extraídos para a construção do modelo e nesta Seção será apresentado a relevância de cada um dos grupos de atributos utilizados no trabalho. A Tabela 4.1 apresenta a correlação de Pearson juntamente com o desvio padrão no uso da predição do efeito de mutações na interação proteína-proteína.

Tabela 4.1: Correlação de Pearson juntamente com o desvio padrão (em $KCal/mol^{-1}$) para cada um dos grupos de atributos nos 4 algoritmos de regressão utilizados neste trabalho.

Grupo	**SVR**	**Árvore M5**	**Processo Gaussiano**	**KNN**
Estrutura do grafo	0,38(1,79)	0,39(1,78)	0,38(1,78)	0,42(1,75)
Classificação dos contatos	0,35(1,81)	0,38(1,77)	0,36(1,73)	0,33(1,82)
Atributos auxiliares	0,20(1,95)	0,21(1,95)	0,24(1,94)	0,20(1,75)
Graph Kernel	0,49(1,60)	0,46(1,61)	0,48(1,60)	0,44(1,61)

A modelagem proposta nesta tese mostra que somente o atributo de *Graph Kernel* tem uma relevância maior que os demais, e, classificação dos contatos e Estrutura do grafo ainda são melhores que os atributos auxiliares que outros trabalhos como [Zhao et al., 2014, Li et al., 2016] as utilizam afim de melhorar o modelo.

Tanto as métricas de estrutura do grafo quanto o *Graph Kernel* são relevantes para o modelo uma vez que sua composição abstrai a relação entre os resíduos da interface em função dos contatos ali formados. Esta configuração apresenta uma melhor representação do efeito da mutação na interação proteína-proteína utilizando apenas dados físico-químicos da estrutura.

4.1.3 Seleção de atributos

O Passo seguinte foi a remoção de atributos por eliminação apresentado na Seção 3.6.4. Para isto os quatro algoritmos selecionaram o melhor conjunto de atributos removendo os que apenas forneciam ruídos ao modelo. A Figura 4.2 apresenta a qualidade do processo de regressão para cada um dos algoritmos utilizados.

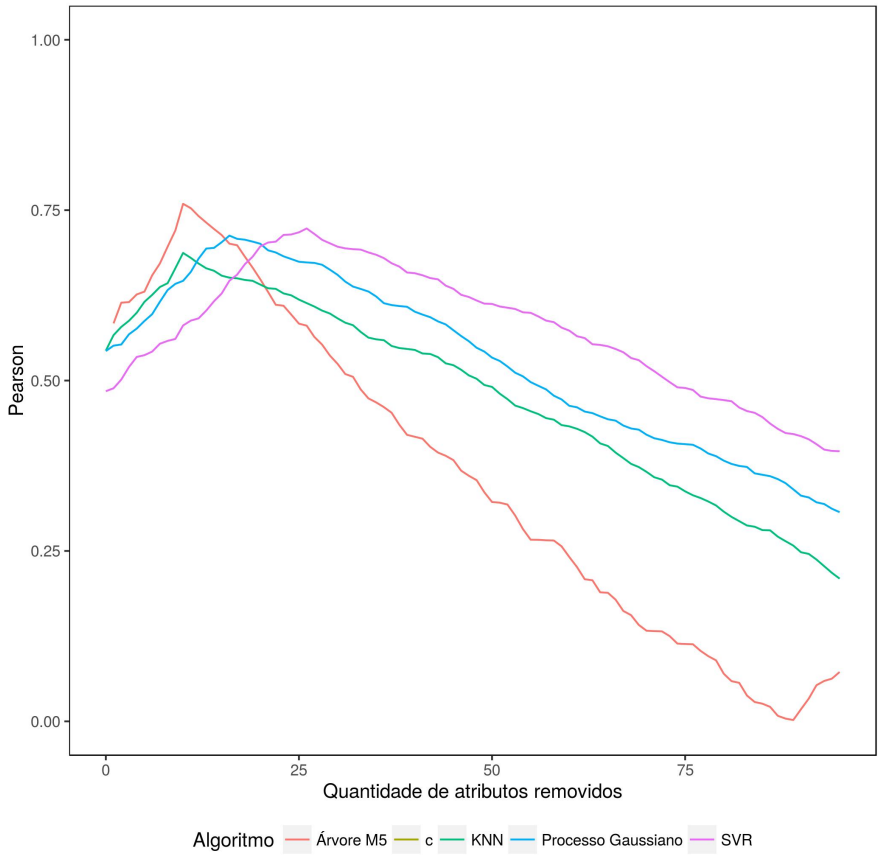


Figura 4.2: Progressão da correlação de Pearson pela quantidade de atributos removidos para cada um dos algoritmos de aprendizado supervisionado.

Como cada algoritmo aborda o problema de uma forma diferente um do outro é possível observar que o melhor conjunto de cada um dos cenários é distinto tanto na qualidade do resultado quanto na quantidade total de atributos. Outro fator importante nesta etapa é que nenhum dos algoritmos sendo utilizados de forma separada nos dá um resultado significativo frente aos demais da literatura. Isto

Tabela 4.2: Quantidade de atributos utilizados em cada algoritmo e seu coeficiente de Pearson para a predição da afinidade do complexo proteico.

Algoritmo	Quantidade de atributos	Pearson	Desvio padrão ($KCal/mol^{-1}$)
KNN	81	0,69	1,53
Árvore M5	81	0,75	1,43
SVR	74	0,72	1,48
Processo Gaussiano	79	0,71	1,53

mostra a importância do uso do *Stacking* para melhorar a qualidade do resultado Final.

A Tabela 4.2 apresenta o melhor resultado obtido por cada um dos algoritmos, o desvio padrão juntamente com a quantidade total de atributos enquanto a Figura 4.4 a dispersão da correlação de pearson por dobra de cada um dos algoritmos. O Apêndice B apresenta com detalhes quais os atributos utilizados por cada algoritmo.

Apesar da árvore M5 possuir a maior correlação de Pearson de todos os 4 algoritmos o mesmo possui a maior dispersão em suas dobras possuindo um cenário onde uma dobra possui o Pearson de 0,42 e outra com 0,92. O SVR também possui uma configuração similar, no qual uma dobra também possui o Pearson de 0,42 e outra com 0,84. Uma outra situação também encontrada é o cenário em que o processo Gaussiano em que um único cenário foi diferente dos demais com a correlação de Pearson de 0,41. Isso mostra que em alguns cenários específicos o próprio conjunto de dados possui situações em que a predição não é eficiente.

Ao avaliar o conjunto de mutações das dobras destes 3 algoritmos foi possível observar que em todas elas que no conjunto de teste constituído de 200 mutações 100 delas o $\Delta\Delta G$ era maior que 5 $KCal/mol^{-1}$. Em outras palavras, o conjunto de treino em momento algum teve como exemplo um valor maior que 5 $KCal/mol^{-1}$ possibilitando o viés aqui apresentado. O Algoritmo KNN passou por este mesmo cenário mas o seu erro foi menor devido a sua heurística, o fato dele considerar como critério a similaridade entre as instâncias é possível observar que o seu resultado se deu pelo fato de que ao escolher os vizinhos mais próximos foram justamente os que possuíam o maior valor da variação da energia livre de ligação no conjunto

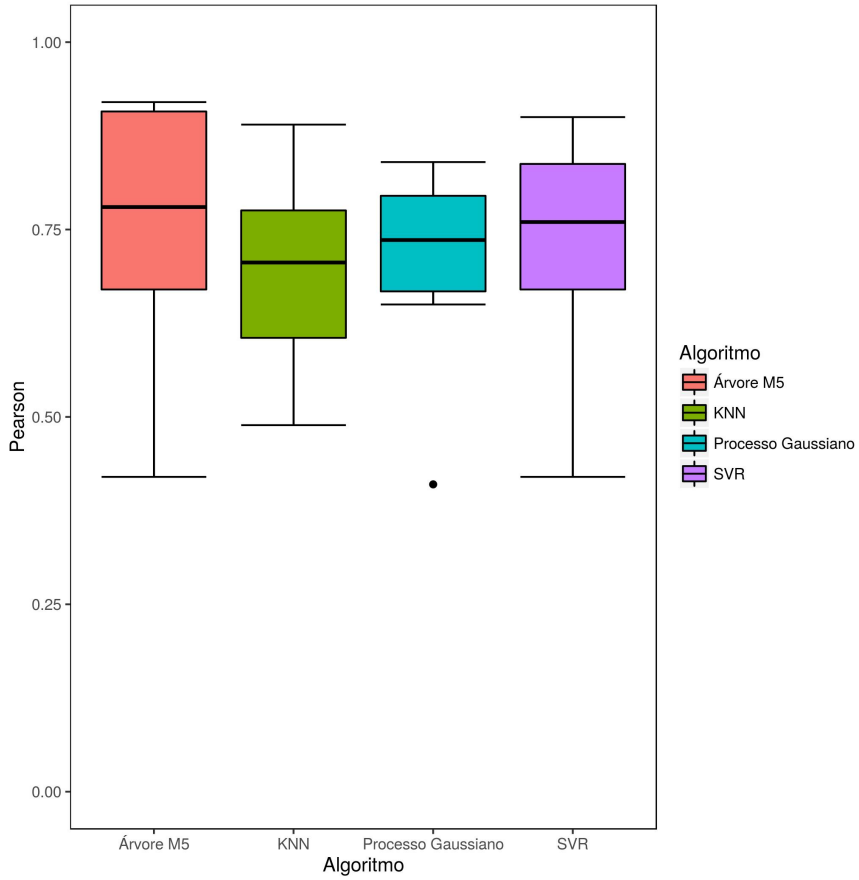


Figura 4.3: Box Plot da correlação de Pearson para as 10 dobras na validação cruzada de cada um dos algoritmos utilizados.

de treino.

Entretanto, nos casos das dobras de teste com a maior correlação de Pearson o intervalo de valores do $\Delta\Delta G$ compreendia entre 0 e 2,5 $KCal/mol^{-1}$ sendo estes os casos de maior frequência na base, e, consequentemente no processo de aprendizagem os algoritmos aqui utilizados foram capaz de modelar de forma mais precisa estas situações.

4.1.4 Redução de dimensionalidade

O uso do SVD se mostrou satisfatório na construção do conjunto de treino. A identificação do posto utilizado no trabalho foi feito de forma que foi testado

todos os postos até encontrar o que proporcionou melhor resultado. O ganho em cada um dos algoritmos é mostrado na Tabela 4.3 juntamente com o p-valor de cada um dos algoritmos.

Tabela 4.3: Tabela com os resultados referentes ao uso do SVD apresentando o ganho no coeficiente de Pearson em relação à Tabela 4.2 juntamente com o teste de significância deste ganho desta correlação.

Algoritmo	Pearson Com SVD	Desvio Padrão $KCal/mol^{-1}$	Ganho no coeficiente	p-valor
KNN	0,73	1,49	0,04	0,01
Árvore M5	0,75	1,43	0,04	0,39
SVR	0,76	1,48	0,04	0,01
Processo Gaussiano	0,74	1,53	0,02	0,06

O uso do SVD se tornou significativo para os algoritmos KNN e SVR pois a melhora foi significativa, mas no processo gaussiano e na árvore M5 o mesmo não aconteceu. Mesmo não sendo um ganho significativo em todos os casos o mesmo foi aplicado nos 4 algoritmos uma vez que em todos eles o Pearson aumentou.

Utilizando os mesmos conjuntos nas dobras em ambos os testes é notório a redução do intervalo interquartil nos resultados dos algoritmos KNN, Processo Gaussiano e SVR. Tanto o Processo Gaussiano quanto a árvore M5 não obtiveram melhora em relação à dobra com a pior correlação de Pearson e nem com a de melhor resultado, entretanto o Processo Gaussiano diminuiu a dispersão dos valores de Pearson com o SVD. Já o SVR ficou com 2 situações fora do seu intervalo interquartil, sendo o menor Pearson referente à dobra já apresentada anteriormente e a segunda com menor valor refere-se à uma dobra com mutações no intervalo de 3 à 5 $KCal/mol^{-1}$ de tal forma que o SVR para estas situações não possui um resultado favorável.

4.1.5 Modelo de Regressão final

A partir dos resultados obtidos de cada um dos 4 algoritmos o processo de resposta apenas é concluindo quando o valor predito pelos 4 primeiros algoritmos são utilizados no Algoritmo de consenso sendo neste caso a *Random Forest Tree* proporcionando um resultado mais preciso do que o uso isolado de cada um dos

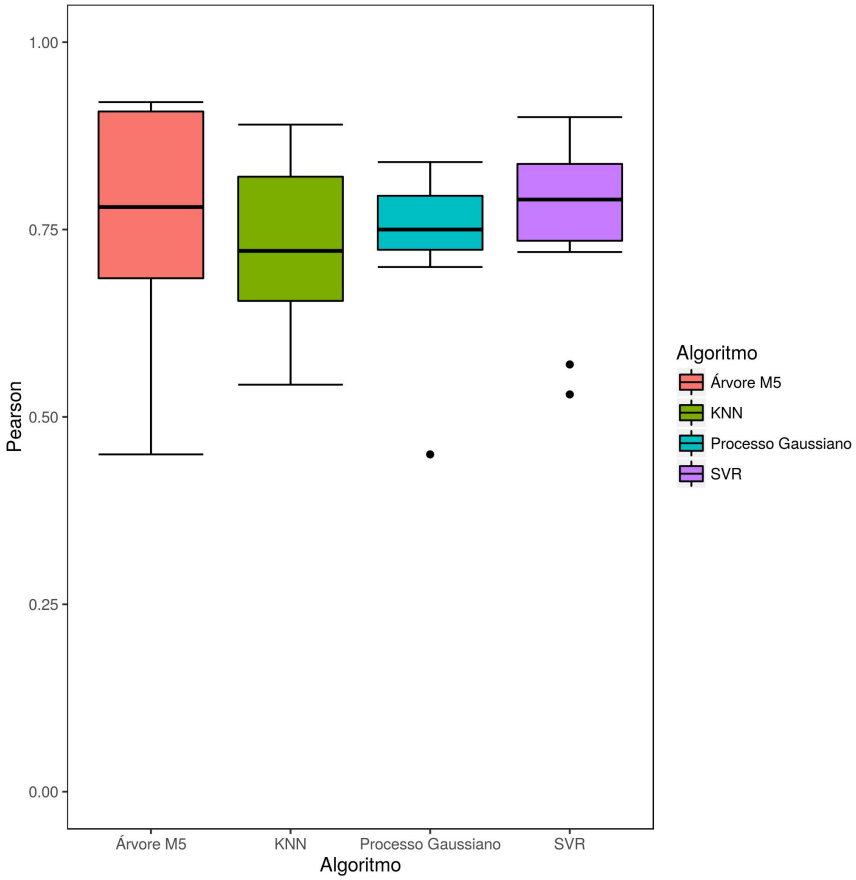


Figura 4.4: Box Plot da correlação de Pearson para as 10 dobras na validação cruzada de cada um dos algoritmos utilizados usando o SVD para redução de dimensionalidade e remoção de ruídos.

resultados. A Figura 4.5 apresenta o gráfico da correlação de Pearson da regressão utilizando *Rando Forest Tree* e agrupando o conjunto predito em função do erro no modelo.

A correlação de Pearson obtida pela validação cruzada de 10 dobras foi de 0,84 com σ=1,39 utilizando todas as 2.007 instâncias da base, e, retirando o conjunto de predições com maior erro o Pearson se eleva para 0,89 com σ=1,32 e ainda o resultado obtêm os Pearsons de 0,91 e 0,94 com desvio padrão 1,29 e 1,27 respectivamente ao retirar os 10% e 20% casos com maior erro.

As mutações encontradas no conjunto de 5% com maior erro em geral são mutações em que a variação da energia livre de ligação não condiz com outras

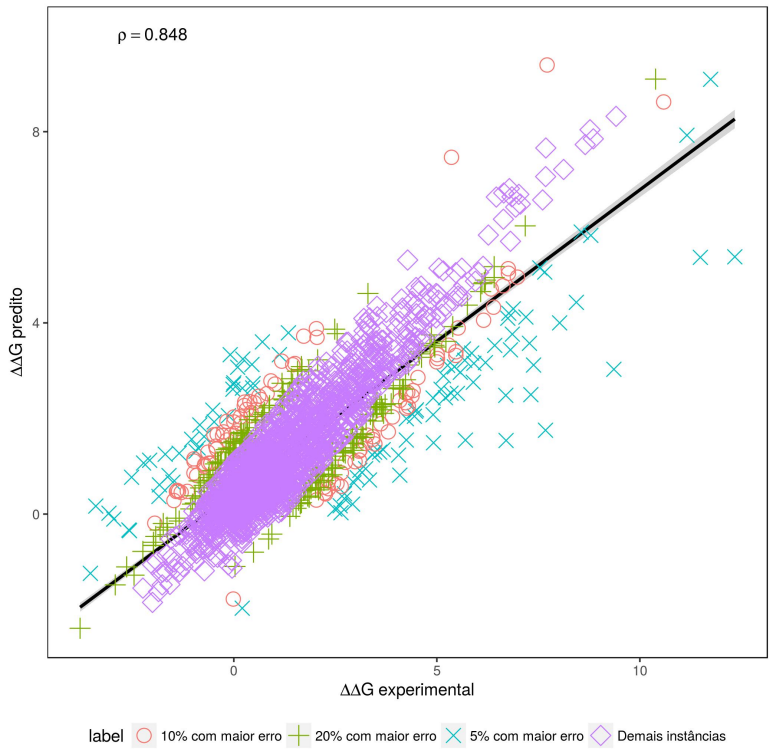


Figura 4.5: Curva de valores singulares de cada um dos algoritmos de regressão.

mutações na mesma posição mas com resíduos diferentes e mesmo assim não foi identificado nenhum padrão nem em relação ao resíduo mutante nem ao resíduo selvagem.

Em contrapartida, o conjunto dos 10% com maior erro possui não somente os casos em que não se encaixam em um grupo específico mas também conjuntos praticamente completos. Das 19 mutações registradas no SKEMPI do PDB 2FTL 18 estão no conjunto de maior erro além de mutações da única mutação dos PDBs 1GL1 e 1HE8 no qual ambos possuem apenas uma única ocorrência no conjunto mostrando que a modelagem da interface destes casos não foi suficientemente eficaz para predizer de forma correta.

4.2 Comparação com outros trabalhos da literatura

Como discutido na Seção 2.2 diferentes ferramentas foram desenvolvidas para a predição da afinidade do complexo proteico utilizando diferentes conjuntos de mutações sendo assim cada trabalho será comprado de forma separada dos demais.

4.2.1 BeAtMuSiC e mCSM

Ambos os trabalhos utilizaram o mesmo conjunto de dados e cenários semelhantes para este fim. A Tabela 4.4 apresenta a relação do coeficiente de Pearson dos 3 trabalhos em diferentes cenários.

Tabela 4.4: Comparação do MutaGraph com os softwares BeAtMuSiC e mCSM nos casos de validação cruzada de 10 dobras,não redundante em posição e proteína.

Modelo	Validação cruzada 10 dobras	Validação cruzada não-redundante em proteína	Validação cruzada não-redundante em posição
MutaGraph	0,84	0.58	0,62
mCSM	0,81	0,58	0,57
BeAtMuSiC	0,40	0,40	0,40

O trabalho proposto nesta tese apresenta um coeficiente de Pearson maior que os demais trabalhos que utilizam as 2.007 mutações singulares disponíveis na base de dados SKEMPI. Em todos os cenários de validação cruzada há significância estatística entre o MutaGraph e as demais ferramentas sendo o p-valor de 0,01, 0,42 e 0,01 para os cenários de validação cruzada de 10 dobras, não-redundante em posição e não-redundante em proteína respectivamente comparado ao mCSM e o mesmo comparativo com o BeAtMuSiC o p-valor é muito próximo de 0 mostrando que as melhoras obtidas nestes cenários pelo modelo proposto têm significância estatística.

4.2.2 MutaBind

A ferramenta MutaBind utiliza em seus testes diferentes conjunto de mutações e validações distintas da proposta pelos outros trabalhos. A Tabela 4.5 apresenta os resultados comparando o MutaGraph com os resultados apresentados por [Li et al., 2016].

Tabela 4.5: Comparação do MutaGraph com os software MutaBind no cenários de validação cruzada de 5 e 4 além do conjunto de mutações apenas com mutações em proteínas inibidoras de protease em dobras com complexos proteícos similares (SKEMPI_MutaBind).

Base	**Validação cruzada**	**MutaGraph**	**MutaBind**	**BeAtMuSiC**
SKEMPI_MutaBind	CV5	**0,68**	0,57	0,39
	CV4	**0,79**	0,68	0,39
SKEMPIpi	CV4	**0,89**	0,76	0,44

Nos testes de validação cruzada com 5 dobras ("CV5") e 4 ("CV4") em que as mutações que utilizam o mesmo complexo e demais complexos que foram considerados idênticos/similares ficaram na mesma dobra, ou seja, mutações onde a região de ligações eram semelhantes ficaram em dobras distintas. Diferente da base testada em 4.2.1 neste caso uma base menor com um conjunto de mutações implica num erro menor justificando o valor de Pearson maior do que o encontrado na Tabela 4.5.

Uma limitação encontrada nos resultados do MutaBind está no intervalo do resultado obtido pelo preditor. Os valores de $\Delta\Delta G$ obtidos vão de -1,5 até 5 $KCal/mol^{-1}$ enquanto o presente trabalho atinge o intervalo maior, de -2,5 até 10 $KCal/mol^{-1}$ tornando-o mais próximo dos efeitos encontrados na base de dados SKEMPI.

4.3 Estudo de Caso: Conjunto de mutações do CAPRI (Rodada 26)

Em [Moretti et al., 2013] apresenta um conjunto de 1.862 mutações de 2 diferentes inibidores de influenza ligados à hemaglutinina. O primeiro conjunto (T55) possui

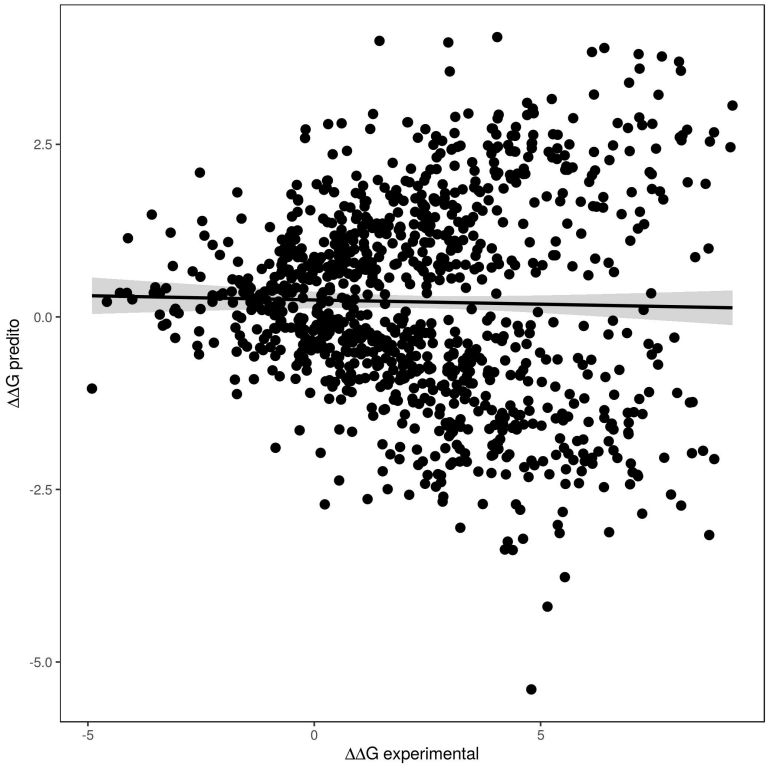


Figura 4.6: Correlação de Pearson para o conjunto de mutações T55 utilizando o modelo construído apenas com o conjunto de mutações com $rho = -0.031$.

1.007 mutações da estrutura do PDB 3R2X enquanto o PDB 4EEF (T56) têm 855 mutações. Nestes testes foi possível comparar o modelo proposto neste trabalho com outros 23 algoritmos utilizados para este mesmo fim.

As Figuras 4.6 e 4.7 apresentam a correlação e Pearson baixa em comparação aos obtidos no modelo SKEMPI. Apesar disso o *score* é o segundo maior obtido para o conjunto T55 mas o menor dos trabalhos correlatos na base T56 4.6.

Tabela 4.6: *Kendall's score* do MutaGraph com demais trabalhos no conjunto T55.

Trabalho	***Kendall's score***
MutaBind	0,41
Mutagraph	0,31
BeAtMuSiC	0,29
FoldX	0,12

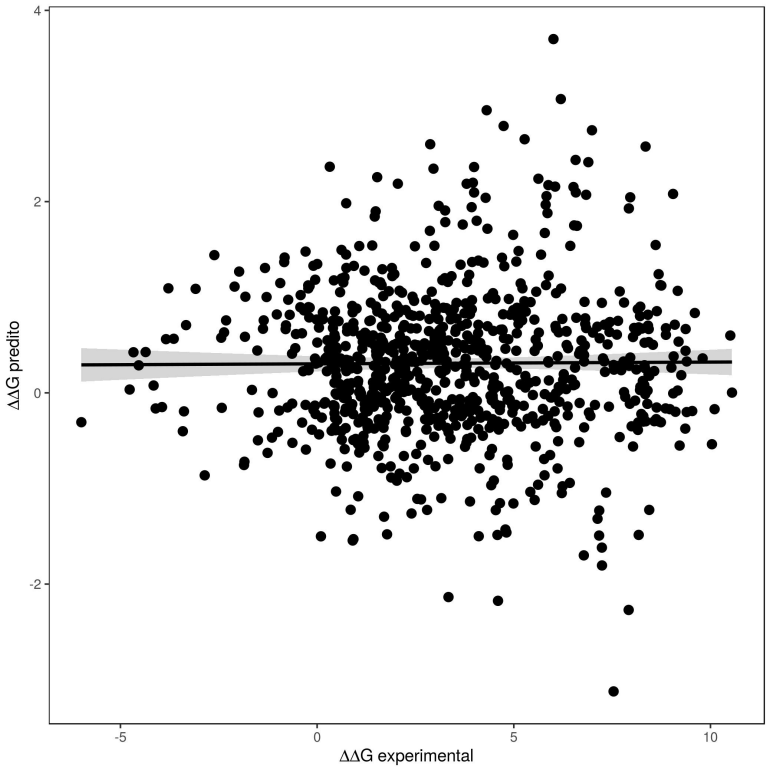


Figura 4.7: Correlação de Pearson para o conjunto de mutações T56 utilizando o modelo construído apenas com o conjunto de mutações $rho = 0.006$.

Tabela 4.7: *Kendall's score* do mutagraph com demais trabalhos no conjunto T56.

Trabalho	***Kendall's score***
MutaBind	0,30
BeAtMuSiC	0,18
FoldX	0,16
Mutagraph	0,13

Este cenário mostra que o modelo proposto neste trabalho teve limitações quanto à modelagem do conjunto T56, e no caso do conjunto T55 o resultado fica abaixo do MutaBind. Apesar do conjunto de atributos possuírem certa semelhança, o uso destes não garantiram um bom resultado (uma vez que a sua contribuição para o resultado também se mostrou baixa).

Um outro fator que influencia na qualidade do resultado de todos os trabalhos é o conjunto de dados utilizado para treino e teste. No caso do uso do SKEMPI

nenhum trabalho atingiu um *score* significativo neste problema, de modo que podemos concluir que é possível existir características nestas mutações e que o SKEMPI não abrange.

Capítulo 5

Web Service

Neste capítulo serão apresentados os recursos desenvolvidos para a visualização dos resultados obtidos a partir dos atributos gerados para a predição do efeito da mutação.

A ferramenta MutaGraph tem por objetivo ser interativa e capaz de visualizar o efeito da mutação mostrando a diferença entre os contatos encontrados na estrutura. A Figura 5.1 apresenta o *layout* da página principal do *webservice*.

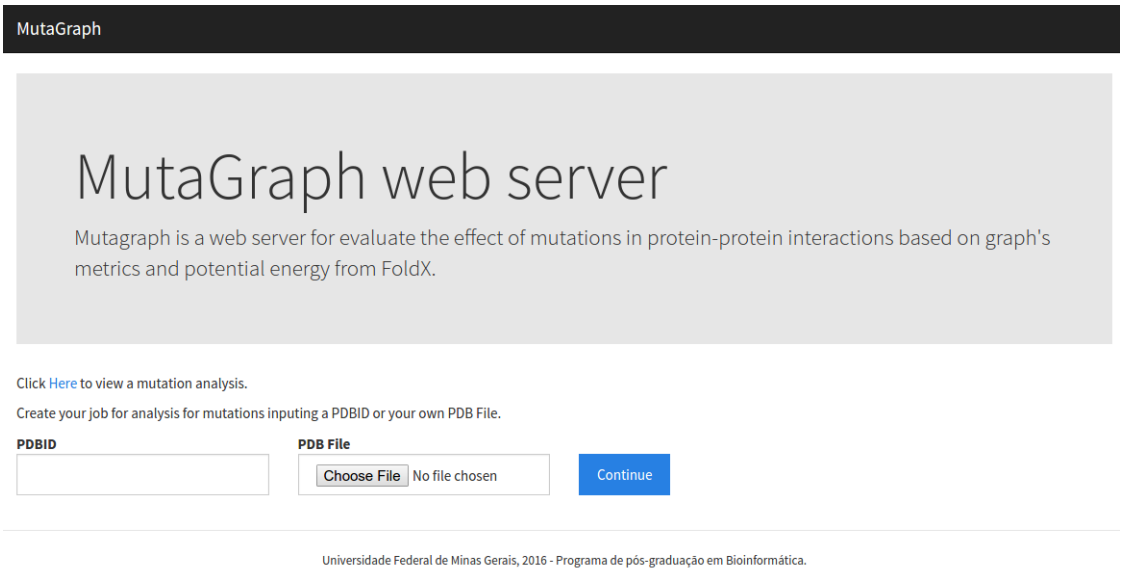


Figura 5.1: Página principal do *webservice*, a partir dela é possível acessar as funções de criar um novo job e consultar o status da que já foram criadas.

5.1 Construção do *Job*

No contexto do serviço disponibilizado na *web*, o *Job* é um conjunto de mutações em que são escalonadas no servidor afim de serem processadas. Devido ao alto custo computacional dos algoritmos utilizados na extração dos atributos. O primeiro passo para a construção do *Job* é informar qual estrutura será analisada. Na página principal do *webservice* (Figura 5.1) é necessário informar o Código PDB (PDBID) ou selecionar um arquivo do computador local com o conteúdo no formato PDB para ser enviado ao servidor.

Quando o arquivo é enviado o passo seguinte é a escolha das posições e os resíduos mutantes. Numa única Job é possível selecionar uma ou mais posições da estrutura para a troca do aminoácido. A Figura 5.2 apresenta tela para a seleção da posição e resíduo mutante sendo que existe ainda a opção de mutação sistemática (no menu *"Mutant"* a opção *"Systematic"*) onde será adicionado à Job 19 mutações referente à troca do resíduo da posição em questão pelos demais.

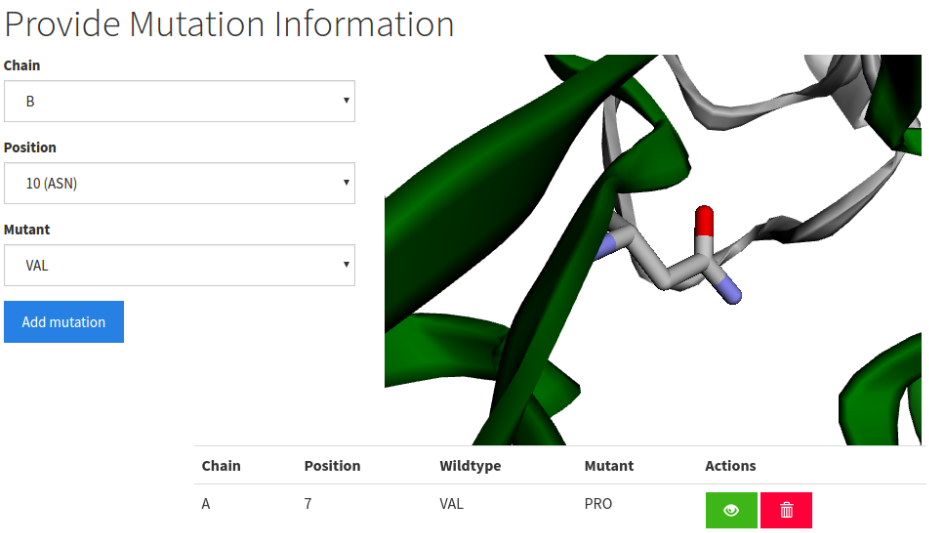


Figura 5.2: Tela para seleção do conjunto de mutações que será processada. Nesta etapa o usuário é capaz de selecionar a cadeia, posição e o resíduo mutante.

Para que o usuário identifique em qual região da estrutura o resíduo será substituído ao lado direito do menu para a escolha da posição em que ocorrerá a mutação existe um visualizador da estrutura da proteína que apresenta o seguinte

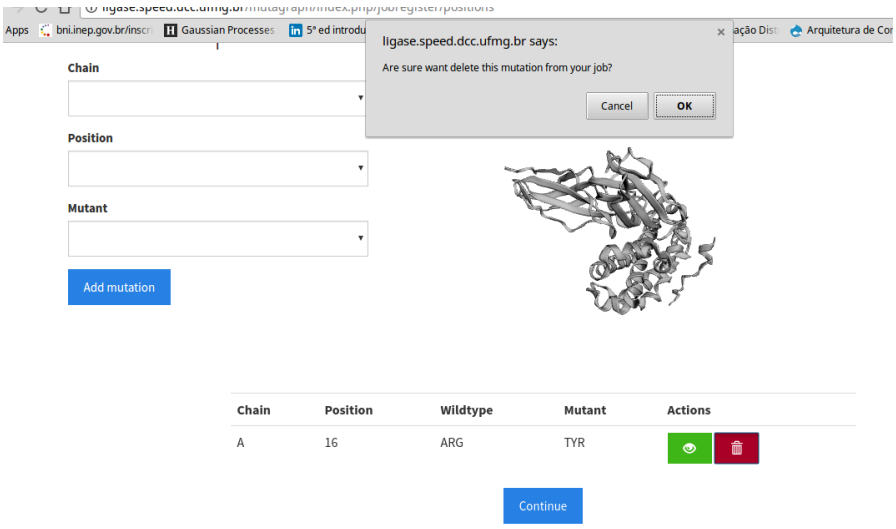


Figura 5.3: Ação de confirmação da remoção de uma determinada mutação da *Job*.

comportamento:

1. Num primeiro momento toda a estrutura é apresentada ao usuário na forma da conformação da estrutura primária em cor cinza.

2. Caso seja escolhida alguma cadeia, a mesma será destacada na cor verde.

3. Quando um resíduo é escolhido o mesmo é destacado na estrutura.

4. Quando uma mutação já está registrada para ser processada, é possível clicar sobre o botão verde localizado na linha da mutação será apresentado na estrutura o resíduo selvagem com o destaque tanto para o resíduo quanto para a sua cadeia.

Quando for necessário o usuário poderá clicar sobre o ícone vermelho localizado na mesma linha da mutação em que deseja remover da *Job*. Como apresentado na Figura 5.3 o usuário irá confirmar sua ação para a sua remoção de mutações compõem o processamento a ser feito no servidor.

Ao apertar o botão de "Continue"no final da página de seleção de mutações o usuário deverá escolher a configuração da interface que será analisada. Para isso será apresentado ao usuário todas as cadeias da estrutura tendo que escolher quais

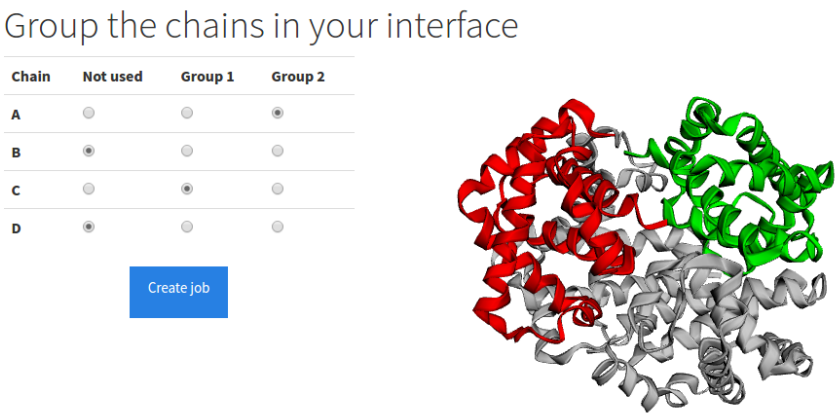


Figura 5.4: Etapa de seleção das cadeias na ferramenta. Todas as cadeias do grupo 1 são destacadas com a cor vermelha enquanto o grupo 2 com a cor verde. Quando uma cadeia não faz parte do grupo então fica com a cor cinza.

as cadeias irão compor os 2 grupos. As cadeias terão suas cores alteradas segundo a configuração feita pelo usuário.

Cada cadeia poderá ser colocada em um dos 2 grupos ou não ser usada na interface. Vale ressaltar que um grupo de cadeias é composto por uma ou várias cadeias e a ferramenta não permite a construção da Job quando não há nenhuma cadeia nos dois grupos ou então um dos grupos ainda é vazio. A Figura 5.4 apresenta esta tela.

A última etapa deste processo então é a construção do job na base de dados e a informação de que a Job foi construída com sucesso. Um código é apresentado na tela para que o usuário seja capaz de consultar a situação da Job no sistema (uma vez que o resultado não é online por questões já discutidas anteriormente) como observado na Figura 5.5. Caso o usuário desejar o próprio sistema enviará um e-mail informando a ele que a tarefa já está concluída, para isto basta o mesmo preencher o campo com o e-mail e esperar a resposta de quando a tarefa concluir. A resposta dada por e-mail pode ser visto na Figura 5.6.

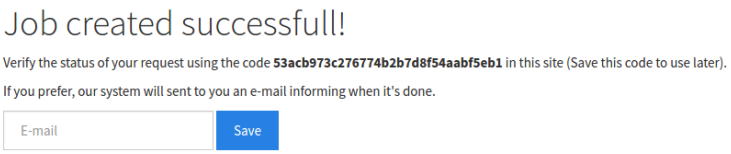


Figura 5.5: Página de conclusão da criação da Job. O usuário pode consultar pelo código o status da *Job* ou também pode registrar seu e-mail para ser notificado quando o processamento terminar.

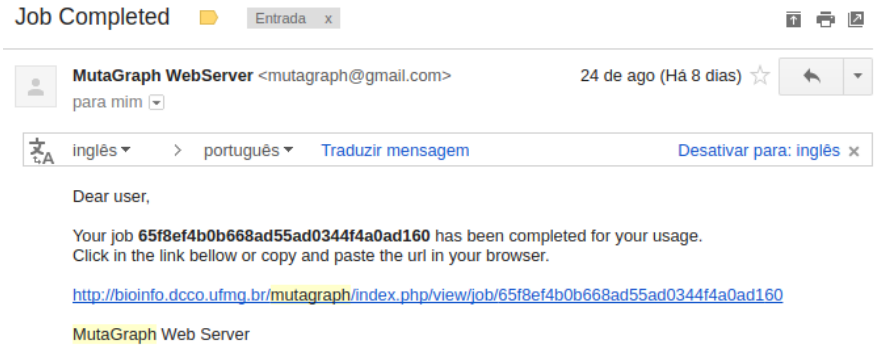


Figura 5.6: Texto com a conclusão do processo de construção e análise da mutação. A partir do próprio e-mail o usuário é capaz de acessar o *webservice* na página de cada Job que requisitou.

5.2 Visualização do efeito da mutação

Com os resultados já processados o usuário é capaz de visualizar e interagir com o grafo de contatos utilizando a ferramenta proposta. Como mostrado na Figura 5.7 o usuário vê simultaneamente tanto o grafo de contatos da interface quanto a estrutura (nela estão a estrutura mutante e a selvagem sobrepostas).

A partir do resultado obtido o usuário poderá interagir afim de entender melhor o efeito que a mutação causou na estrutura. Para isto, a ferramenta disponibiliza um conjunto de possibilidades de manuseio.

A nomenclatura da mutação possui o padrão <PDBID><CADEIA>_<RESÍDUO SELVAGEM><POSIÇÃO><RESÍDUO MUTANTE>(<GRUPO1 >_<GRUPO2>). Este formato é utilizado pois a literatura utiliza uma formatação semelhante para catalogar além da energia livre

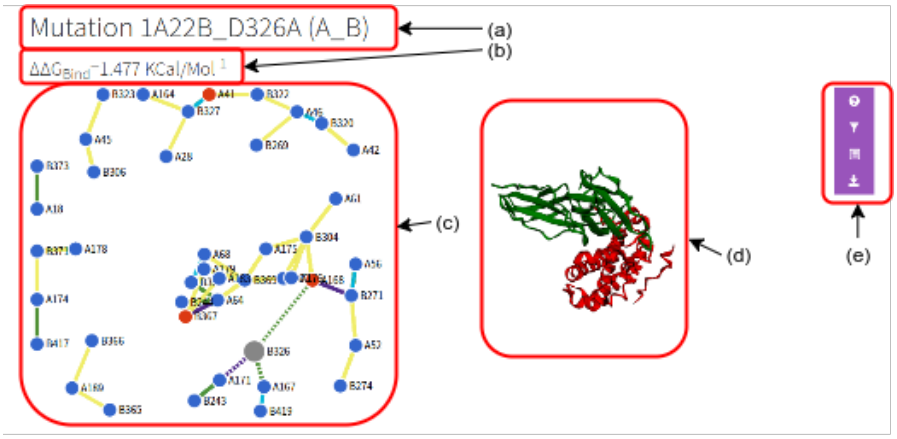


Figura 5.7: Página com as informações do efeito da mutação onde (a) é a nomenclatura utilizada para a mutação, (b) o valor da variação de energia livre de ligação predita pela ferramenta, (c) a estrutura de visualização do grafo de contatos, (d) a estrutura da proteína do mutante e selvagem sobrepostos e (d) o menu de contexto.

de ligação é apresentada com a sua respectiva unidade de medida ($KCal/mol^{-1}$) uma vez que a base utilizada neste trabalho nos fornece os dados suficientes para computar neste formato, como discutido na Seção 3.3.

O grafo de contatos é constituído por vértices que representam os resíduos enquanto as arestas representam os contatos. Cada aresta possui uma cor e um formato distinto afim de diferenciá-las. Quanto à diferença da cor, o mesmo indica qual é o tipo de contato até existe enquanto a forma da aresta indica se a mesma existe apenas no selvagem (quando esta é contínua) ou se é exclusiva da estrutura selvagem (quando a aresta for pontilhada). A cor indica o tipo de contato computado. Apresentado na Seção 3.4, as cores podem ser consultadas no menu de contexto (primeira opção) tendo esta a legenda de todos os elementos da visualização, a Figura 5.8 apresenta o esquema de legendas utilizado no servidor *web*.

E por fim, ainda é possível fazer o download de um arquivo compactado contendo todo o conteúdo gerado pela ferramenta. O último item do menu de contexto permite ao usuário fazer o download de um arquivo compactado no qual possui os seguintes arquivos:

1. Arquivo PDB original (o que foi enviado para o usuário ou o PDB recuperado

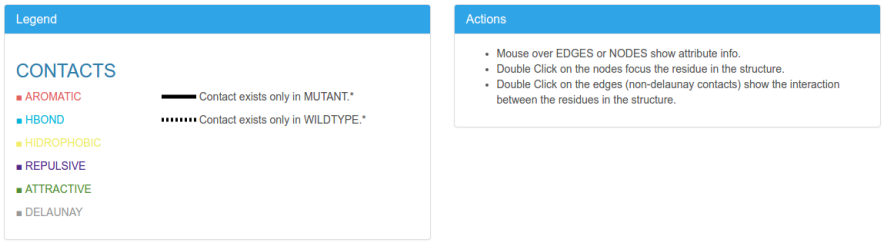


Figura 5.8: Legenda da ferramenta contendo a descrição das cores, formas das arestas e e vértices além das possíveis ações existentes para interação.

do site http://www.rcsb.org/)

2. Arquivo PDB do mutante e selvagem gerado.

3. Arquivo CSV contendo os dados dos contatos que existem apenas ou no selvagem ou no mutante.

4. Arquivo CSV contendo os dados dos atributos computados.

Dessa forma, o usuário poderá interagir com os resultados e ser capaz de visualizar como a mutação alterou os contatos da estrutura. Além disso, o próprio usuário será capaz de avaliar a qualidade do resultado e identificar se existe coerência entre o resultado obtido e as alterações identificadas na interface pela ferramenta.

Capítulo 6

Conclusões e trabalhos futuros

O presente trabalho apresenta uma nova abordagem para o problema de predição de interação proteína-proteína utilizando métricas de grafo e *Graph Kernel*. Apesar de outros trabalhos utilizarem atributos de grafo para tal fim, a Correlação de Pearson obtida pelo método de aprendizado neste trabalho é maior que os demais da literatura uma vez que utiliza a técnica de *Stacking* para melhorar o resultado fazendo o consenso da predição de 4 algoritmos de regressão.

Do conjunto de atributos extraídos o *Graph Kernel* obteve a melhor correlação em todos os 4 algoritmos testados em relação aos demais grupos. Métricas de grafo também se mostraram mais eficientes que o uso de atributos físico-químicos da interface, isto se dá ao fato de que a configuração topológica da rede interações entre proteínas se mostrou relevante para o problema.

Mesmo assim, o modelo apresentou algumas limitações. A primeira delas foi na questão da predição de valores maiores que 5 $KCal/mol^{-1}$ onde nem mesmo a remoção de ruídos com SVD foi suficiente para melhorar a predição nestes casos. No caso da predição do efeito de mutações na interação proteína-proteína da hemaglutinina o modelo teve o pior resultado no conjunto T56 enquanto no T55 o MutaBind obteve um melhor resultado.

A comparação feita com os demais trabalhos da literatura mostraram resultados significantes em relação a eles.

Para um melhor entendimento do efeito da mutação na interface, também foi desenvolvido uma ferramenta *web* de forma que o usuário seja capaz de interagir

e visualizar o efeito da mutação em relação na interação no complexo. Sendo o objetivo da ferramenta *web* proporcionar ao usuário a capacidade de extrair informações da mutação o mesmo é feito de forma a representar a diferença de contatos entre a estrutura mutante e selvagem.

A predição do efeito da mutação na interação proteína-proteína ainda é um problema recente e em aberto na Bioinformática e o uso de métodos computacionais para larga escala é o mais viável comparado aos métodos experimentais.

6.1 Trabalhos futuros

Durante o processo de desenvolvimento da estrutura da metodologia algumas tarefas foram levantadas afim de melhorar o modelo a ser elaborado pode ser o uso de técnicas para melhorar o resultado predito, como por exemplo, utilizar algoritmos genéticos para refinar os parâmetros de alguns algoritmos de aprendizado (como por exemplo o SVR).

Com o uso da ferramenta e o *feedback* de usuários também será possível melhorar o modelo de visualização, isto poderá tornar a ferramenta mais intuitivamente de modo que o efeito da mutação tanto nas mudanças das interações que ocorrem do selvagem para o mutante quanto informações que não são comumente vistos.

Ainda por fim aplicar o trabalho proposto em bases de SNV[1] afim de identificar o efeito de mutações em variantes uma vez que estas podem estar relacionadas ao câncer [Vázquez et al., 2015].

[1] *Single Nucleotide Variant*

Anexo A

Resultado da predição

Abaixo encontra-se a lista completa de cada mutação da base dados SKEMPI. A coluna código da mutação corresponde as informações da mutação no seguinte formato <CÓDIGO PDB><CADEIA>_<RESÍDUO SELVAGEM><POSIÇÃO><RESÍDUO MUTANTE>. O Código PDB são os 4 primeiros caracteres enquanto a cadeia é o quinto. Os resíduos selvagens e mutantes são a primiera e última letra respectivamente após (_) sendo o valor entre eles a posição.

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1A22A_C182A	1,009	1,607	0,598
1A22B_C308A	0,000	1,843	1,843
1A22B_C322A	0,000	1,393	1,393
1A22A_D171A	0,790	1,436	0,646
1A22A_D26A	-0,211	0,003	0,214
1A22B_D364A	1,541	1,706	0,165
1A22B_D405A	-0,046	2,519	2,565
1A22B_D326A	0,991	1,477	0,486
1A22B_D332A	1,085	2,828	1,743
1A22A_E174A	-0,924	0,711	1,635
1A22A_E186A	-0,012	1,072	1,084
1A22A_E56A	0,410	0,225	0,185
1A22A_E65A	-0,473	0,908	1,381
1A22A_E66A	0,430	0,866	0,436
1A22B_E242A	0,705	1,604	0,899
1A22B_E373A	0,083	2,425	2,342
1A22B_E244A	1,380	1,310	0,07
1A22B_E375A	-0,079	2,278	2,357
1A22B_E380A	0,199	1,457	1,258
1A22B_E407A	-0,079	0,453	0,532
1A22B_E409A	-0,062	0,893	0,955
1A22B_E424A	0,108	0,108	0
1A22B_E275A	-0,091	0,858	0,949
1A22B_E279A	-0,079	0,968	1,047
1A22B_E282A	0,782	2,267	1,485
1A22B_E291A	0,178	2,430	2,252
1A22B_E320A	-0,191	2,347	2,538
1A22B_E327A	0,793	1,447	0,654
1A22A_F10A	1,039	0,194	0,845
1A22A_F176A	0,410	2,631	2,221
1A22A_F191A	0,191	0,570	0,379
1A22A_F25A	-0,439	0,957	1,396
1A22A_F54A	0,869	0,077	0,792
1A22B_F296S	-0,041	0,512	0,553
1A22B_F300A	-0,015	2,331	2,346
1A22A_G187A	0,340	1,118	0,778
1A22A_H18A	-0,486	1,597	2,083
1A22A_H21A	0,155	1,632	1,477
1A22A_I179A	0,805	0,366	0,439

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1A22A_I4A	0,408	2,938	2,53
1A22A_I58A	1,637	1,588	0,049
1A22B_I365A	2,163	2,129	0,034
1A22B_I303A	1,279	2,878	1,599
1A22B_I305A	1,356	2,004	0,648
1A22B_I328A	0,588	1,387	0,799
1A22A_K168A	-0,155	0,923	1,078
1A22A_K172A	2,013	0,453	1,56
1A22A_K70A	0,520	2,498	1,978
1A22B_K367A	-0,009	2,128	2,137
1A22B_K379A	0,000	0,147	0,147
1A22B_K403A	0,240	0,353	0,113
1A22B_K406A	0,199	1,322	1,123
1A22B_K415A	0,642	2,552	1,91
1A22B_K281A	0,144	2,796	2,652
1A22B_K310A	0,043	0,321	0,278
1A22B_K321A	0,090	2,218	2,128
1A22A_L15A	0,150	1,126	0,976
1A22A_L45A	1,223	1,978	0,755
1A22A_L6A	0,599	0,509	0,09
1A22A_L73A	-0,200	2,612	2,812
1A22A_L9A	-0,040	2,535	2,575
1A22A_M14A	0,000	1,836	1,836
1A22A_N12A	0,100	1,576	1,476
1A22A_N63A	0,314	2,372	2,058
1A22B_N418A	0,298	1,130	0,832
1A22B_N272A	0,011	1,537	1,526
1A22B_N297A	-0,255	1,294	1,549
1A22A_P48A	0,410	1,858	1,448
1A22A_P59A	0,380	2,409	2,029
1A22A_P5A	0,430	0,471	0,041
1A22A_P61A	1,208	1,127	0,081
1A22B_P306A	2,963	1,221	1,742
1A22B_P241A	0,259	2,724	2,465
1A22A_Q181A	0,270	0,956	0,686
1A22A_Q22A	-0,220	2,609	2,829
1A22A_Q29A	-0,588	1,178	1,766
1A22A_Q46A	0,108	1,064	0,956
1A22A_Q68A	0,588	1,783	1,195

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1A22A_Q69A	-0,050	2,725	2,775
1A22B_Q366A	0,009	2,277	2,268
1A22B_Q416A	0,895	0,207	0,688
1A22B_Q274A	-0,193	0,206	0,399
1A22B_Q278A	-0,410	0,409	0,819
1A22B_Q330A	-0,030	1,425	1,455
1A22A_R167A	0,278	2,901	2,623
1A22A_R178A	2,423	1,228	1,195
1A22A_R16A	0,238	0,438	0,2
1A22A_R183A	0,542	1,507	0,965
1A22A_R19A	0,051	0,221	0,17
1A22A_R64A	1,641	2,583	0,942
1A22A_R8A	0,200	0,252	0,052
1A22B_R243A	1,530	2,511	0,981
1A22B_R243L	0,521	2,799	2,278
1A22B_R243M	0,999	0,464	0,535
1A22B_R411A	0,056	2,920	2,864
1A22B_R413A	-0,190	1,207	1,397
1A22B_R417A	0,268	0,277	0,009
1A22B_R270A	0,553	0,763	0,21
1A22B_R271A	0,567	2,986	2,419
1A22B_R239A	0,269	0,639	0,37
1A22A_S184A	-0,050	0,238	0,288
1A22A_S188A	-0,200	2,274	2,474
1A22A_S51A	0,348	1,655	1,307
1A22A_S55A	0,110	1,454	1,344
1A22A_S57A	0,200	2,303	2,103
1A22A_S62A	0,155	1,152	0,997
1A22A_S71A	0,408	1,853	1,445
1A22A_S7A	0,340	2,809	2,469
1A22B_S247A	-0,015	2,485	2,5
1A22B_S419A	0,034	1,000	0,966
1A22B_S298A	-0,161	0,542	0,703
1A22B_S299A	-0,506	2,713	3,219
1A22B_S302A	0,028	2,936	2,908
1A22B_S324A	0,237	0,543	0,306
1A22A_T175A	1,905	2,907	1,002
1A22A_T3A	-0,050	0,907	0,957
1A22B_T394A	0,202	1,812	1,61
1A22B_T395A	-0,097	0,341	0,438
1A22B_T251A	0,070	2,269	2,199
1A22B_T273A	-0,077	1,319	1,396
1A22B_T277A	-0,025	2,791	2,816
1A22B_T301A	1,423	0,996	0,427
1A22A_V180A	0,000	2,608	2,608
1A22A_V185A	0,869	1,053	0,184
1A22B_V371A	-0,658	0,441	1,099
1A22B_V325A	1,141	2,046	0,905
1A22B_V329A	-0,046	2,905	2,951
1A22B_W276A	0,558	2,388	1,83
1A22B_W280A	0,050	0,224	0,174
1A22B_W304F	2,782	2,956	0,174
1A22A_Y164A	0,348	2,110	1,762
1A22A_Y42A	0,199	1,680	1,481
1A22B_Y295A	0,199	2,069	1,87
1A4YA_D435A	3,482	2,307	1,175
1A4YA_E287A	0,101	2,326	2,225
1A4YA_E344A	0,178	1,709	1,531
1A4YA_E401A	0,883	1,567	0,684
1A4YB_E108A	-0,322	2,707	3,029
1A4YB_H114A	0,656	2,837	2,181
1A4YB_H13A	-0,296	0,063	0,359
1A4YB_H84A	0,170	0,803	0,633
1A4YB_H8A	0,903	0,825	0,078
1A4YA_I459A	0,679	1,716	1,037
1A4YA_K320A	-0,310	0,105	0,415
1A4YB_K40G	3,233	1,377	1,856
1A4YB_K40Q	4,248	0,569	3,679
1A4YB_N68A	0,118	1,713	1,595
1A4YB_Q12A	0,300	0,861	0,561
1A4YA_R457A	-0,224	2,167	2,391
1A4YB_R31A	0,250	2,249	1,999
1A4YB_R32A	0,909	1,743	0,834
1A4YB_R33A	0,327	1,970	1,643
1A4YB_R5A	2,307	1,620	0,687
1A4YB_R66A	0,203	2,188	1,985
1A4YB_R70A	-0,232	0,084	0,316
1A4YA_S289A	0,042	2,853	2,811

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1A4YA_W261A	0,101	2,015	1,914	1BRSA_E73Q	1,450	1,640	0,19
1A4YA_W263A	1,170	1,514	0,344	1BRSA_E73S	3,007	1,968	1,039
1A4YA_W318A	1,499	1,526	0,027	1BRSA_E73W	1,655	1,920	0,265
1A4YA_W375A	1,034	0,087	0,947	1BRSA_E73Y	2,411	0,694	1,717
1A4YB_W89A	0,240	1,859	1,619	1BRSD_E76A	1,093	0,712	0,381
1A4YA_Y434A	3,258	0,035	3,223	1BRSD_E80A	0,476	1,478	1,002
1A4YA_Y434F	0,557	0,595	0,038	1BRSA_H102A	6,340	2,252	4,088
1A4YA_Y437A	0,835	2,869	2,034	1BRSA_H102D	4,546	0,827	3,719
1A4YA_Y437F	0,245	0,564	0,319	1BRSA_H102G	6,813	0,381	6,432
1AHWC_D178A	-0,484	0,210	0,694	1BRSA_H102L	7,658	2,421	5,237
1AHWC_L176A	0,986	2,357	1,371	1BRSA_H102Q	4,546	1,962	2,584
1AHWC_N199A	1,078	2,156	1,078	1BRSA_K27A	5,120	2,546	2,574
1AHWC_T167A	-0,074	0,710	0,784	1BRSA_N58A	3,088	0,845	2,243
1AHWC_T170A	1,105	0,079	1,026	1BRSA_R59A	4,983	2,588	2,395
1AHWC_T197A	1,345	1,269	0,076	1BRSA_R59K	2,485	1,404	1,081
1AHWC_V198A	-0,314	0,105	0,419	1BRSA_R83Q	5,414	2,256	3,158
1AHWC_Y157A	-1,888	2,721	4,609	1BRSA_R87A	5,752	1,066	4,686
1AK4D_A488G	4,015	1,287	2,728	1BRSD_T42A	1,856	1,564	0,292
1AK4D_A488V	2,132	1,272	0,86	1BRSA_W35F	1,259	1,623	0,364
1AK4D_A492G	1,943	2,498	0,555	1BRSD_W38F	1,641	0,424	1,217
1AK4D_A492V	1,721	2,674	0,953	1BRSD_W44F	0,056	0,248	0,192
1AK4D_G489A	3,438	2,876	0,562	1BRSD_Y29A	3,467	2,968	0,499
1AK4D_G489V	4,390	0,360	4,03	1BRSD_Y29F	-0,132	2,096	2,228
1AK4D_H487A	2,372	2,169	0,203	1CHOI_A15C	-0,326	1,576	1,902
1AK4D_H487Q	2,335	0,768	1,567	1CHOI_A15D	0,000	1,593	1,593
1AK4D_H487R	3,007	0,333	2,674	1CHOI_A15E	0,225	0,238	0,013
1AK4D_I491A	1,603	2,359	0,756	1CHOI_A15F	-0,735	2,412	3,147
1AK4D_I491V	1,363	1,797	0,434	1CHOI_A15G	0,066	2,345	2,279
1AK4D_P485A	2,447	1,908	0,539	1CHOI_A15H	-0,326	1,489	1,815
1AK4D_P490A	3,533	1,150	2,383	1CHOI_A15I	0,655	0,409	0,246
1AK4D_P493A	2,045	0,026	2,019	1CHOI_A15K	2,496	0,580	1,916
1AK4D_V486A	2,353	2,679	0,326	1CHOI_A15L	0,141	1,674	1,533
1BRSA_D54A	-0,528	1,317	1,845	1CHOI_A15M	0,141	0,628	0,487
1BRSD_D35A	4,281	1,548	2,733	1CHOI_A15N	-0,289	1,121	1,41
1BRSD_D39A	6,786	0,791	5,995	1CHOI_A15P	3,105	0,717	2,388
1BRSA_E60A	0,094	2,308	2,214	1CHOI_A15Q	0,442	0,224	0,218
1BRSA_E73A	2,350	0,168	2,182	1CHOI_A15R	0,468	1,969	1,501
1BRSA_E73C	2,526	2,916	0,39	1CHOI_A15S	0,066	0,173	0,107
1BRSA_E73F	2,233	2,078	0,155	1CHOI_A15T	0,968	0,956	0,012

Código da mutação	ΔΔG Experimental	ΔΔG Predito		Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1CHOI_A15V	0,937	1,264		1CHOI_G32W	1,055	2,258	1,203
1CHOI_A15W	-1,923	0,452		1CHOI_G32Y	-0,982	2,143	3,125
1CHOI_A15Y	-0,875	1,467		1CHOI_K13A	0,181	0,942	0,761
1CHOI_E19A	2,331	2,979		1CHOI_K13C	0,503	1,281	0,778
1CHOI_E19C	2,725	1,494		1CHOI_K13D	-0,510	0,577	1,087
1CHOI_E19D	0,881	1,534		1CHOI_K13E	0,181	2,041	1,86
1CHOI_E19F	3,619	0,186		1CHOI_K13F	1,092	2,475	1,383
1CHOI_E19G	3,619	2,329		1CHOI_K13G	1,820	0,254	1,566
1CHOI_E19H	2,087	2,085	0,002000000000000002	1CHOI_K13H	0,141	2,845	2,704
1CHOI_E19I	2,639	0,440		1CHOI_K13I	1,543	1,461	0,082
1CHOI_E19K	3,591	1,946		1CHOI_K13L	0,380	1,769	1,389
1CHOI_E19L	2,866	2,409		1CHOI_K13M	0,426	1,464	1,038
1CHOI_E19M	2,695	0,205		1CHOI_K13N	-0,361	2,013	2,374
1CHOI_E19N	2,331	2,801		1CHOI_K13P	2,074	2,672	0,598
1CHOI_E19P	2,588	1,390		1CHOI_K13Q	-0,572	1,323	1,895
1CHOI_E19Q	1,333	1,545		1CHOI_K13R	-0,377	1,597	1,974
1CHOI_E19R	3,678	0,555		1CHOI_K13S	-0,426	1,016	1,442
1CHOI_E19S	2,791	1,141		1CHOI_K13T	0,380	0,288	0,092
1CHOI_E19T	2,906	1,040		1CHOI_K13V	1,201	1,391	0,19
1CHOI_E19V	2,437	1,776		1CHOI_K13W	1,018	0,092	0,926
1CHOI_E19W	3,156	2,760		1CHOI_K13Y	1,074	2,287	1,213
1CHOI_E19Y	2,666	2,472		1CHOI_L18A	4,844	2,158	2,686
1CHOI_G32A	-0,931	1,080		1CHOI_L18C	2,639	0,931	1,708
1CHOI_G32C	2,541	0,985		1CHOI_L18D	7,193	1,181	6,012
1CHOI_G32D	3,105	1,685		1CHOI_L18E	6,748	0,112	6,636
1CHOI_G32E	2,613	2,471		1CHOI_L18F	-1,500	0,017	1,517
1CHOI_G32F	0,102	1,299		1CHOI_L18G	6,079	2,750	3,329
1CHOI_G32H	-0,982	2,081		1CHOI_L18H	2,997	1,700	1,297
1CHOI_G32I	2,365	0,507		1CHOI_L18I	4,474	0,247	4,227
1CHOI_G32K	1,587	2,295		1CHOI_L18K	4,368	2,136	2,232
1CHOI_G32L	1,749	0,953		1CHOI_L18M	0,352	0,937	0,585
1CHOI_G32M	1,635	0,878		1CHOI_L18N	3,353	1,280	2,073
1CHOI_G32N	0,952	1,531		1CHOI_L18P	8,792	0,587	8,205
1CHOI_G32P	0,324	2,024		1CHOI_L18Q	3,252	2,094	1,158
1CHOI_G32Q	2,381	2,525		1CHOI_L18R	4,002	2,709	1,293
1CHOI_G32R	1,743	0,555		1CHOI_L18S	4,981	1,265	3,716
1CHOI_G32S	-0,137	1,199		1CHOI_L18T	4,480	1,196	3,284
1CHOI_G32T	2,074	1,329		1CHOI_L18V	4,269	2,184	2,085
1CHOI_G32V	0,646	0,049		1CHOI_L18W	-1,689	1,335	3,024

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1CHOI_L18Y	-2,223	0,120	2,343
1CHOI_N28S	0,000	2,598	2,598
1CHOI_N36A	-1,363	0,630	1,993
1CHOI_N36C	-0,112	1,354	1,466
1CHOI_N36D	-1,039	1,881	2,92
1CHOI_N36E	2,331	2,345	0,014
1CHOI_N36F	2,365	0,079	2,286
1CHOI_N36G	-1,393	2,361	3,754
1CHOI_N36H	-0,735	2,570	3,305
1CHOI_N36I	-0,030	2,764	2,794
1CHOI_N36K	2,331	0,113	2,218
1CHOI_N36L	0,616	1,271	0,655
1CHOI_N36M	-0,208	0,016	0,224
1CHOI_N36P	-1,832	2,265	4,097
1CHOI_N36Q	-0,185	2,280	2,465
1CHOI_N36R	1,587	1,603	0,016
1CHOI_N36S	-1,475	1,176	2,651
1CHOI_N36T	-1,297	1,073	2,37
1CHOI_N36V	0,033	1,835	1,802
1CHOI_N36W	2,176	2,955	0,779
1CHOI_N36Y	2,563	0,271	2,292
1CHOI_P14A	0,380	1,491	1,111
1CHOI_P14C	-0,982	0,643	1,625
1CHOI_P14D	-0,850	1,742	2,592
1CHOI_P14E	-0,719	0,779	1,498
1CHOI_P14F	-0,307	1,670	1,977
1CHOI_P14G	0,636	0,011	0,625
1CHOI_P14H	-0,112	1,047	1,159
1CHOI_P14I	-0,510	2,201	2,711
1CHOI_P14K	0,386	1,866	1,48
1CHOI_P14L	-0,524	2,724	3,248
1CHOI_P14M	-0,827	2,023	2,85
1CHOI_P14N	-0,779	1,942	2,721
1CHOI_P14Q	-0,727	0,643	1,37
1CHOI_P14R	0,398	0,388	0,01
1CHOI_P14S	-0,440	0,012	0,452
1CHOI_P14T	-0,549	1,895	2,444
1CHOI_P14V	-0,549	0,438	0,987
1CHOI_P14W	-0,208	2,506	2,714
1CHOI_P14Y	-0,395	2,215	2,61
1CHOI_R21A	3,188	0,558	2,63
1CHOI_R21C	3,527	2,146	1,381
1CHOI_R21D	3,779	0,055	3,724
1CHOI_R21E	3,504	2,837	0,667
1CHOI_R21F	2,950	2,020	0,93
1CHOI_R21G	3,482	1,590	1,892
1CHOI_R21H	2,791	1,890	0,901
1CHOI_R21I	2,107	1,677	0,43
1CHOI_R21K	0,662	2,719	2,057
1CHOI_R21L	2,827	2,562	0,265
1CHOI_R21M	2,613	1,652	0,961
1CHOI_R21N	3,527	0,574	2,953
1CHOI_R21P	7,281	0,915	6,366
1CHOI_R21Q	2,476	2,852	0,376
1CHOI_R21S	3,183	2,515	0,668
1CHOI_R21T	2,666	1,020	1,646
1CHOI_R21V	2,064	0,553	1,511
1CHOI_R21W	2,437	2,706	0,269
1CHOI_R21Y	2,997	2,870	0,127
1CHOI_T17A	4,234	0,234	4
1CHOI_T17C	3,325	1,557	1,768
1CHOI_T17D	6,813	1,084	5,729
1CHOI_T17E	5,655	0,751	4,904
1CHOI_T17F	4,228	0,502	3,726
1CHOI_T17G	5,289	0,259	5,03
1CHOI_T17H	3,648	2,787	0,861
1CHOI_T17I	3,648	1,157	2,491
1CHOI_T17K	4,088	0,425	3,663
1CHOI_T17L	4,412	1,655	2,757
1CHOI_T17M	3,951	0,998	2,953
1CHOI_T17N	4,313	0,024	4,289
1CHOI_T17P	4,228	0,127	4,101
1CHOI_T17Q	3,678	0,565	3,113
1CHOI_T17R	3,880	2,444	1,436
1CHOI_T17S	2,348	1,524	0,824
1CHOI_T17V	3,424	0,288	3,136
1CHOI_T17W	4,154	2,867	1,287
1CHOI_T17Y	4,468	1,472	2,996

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1CHOI_Y20A	2,541	1,758	0,783	1DANT_F76A	1,104	2,145	1,041
1CHOI_Y20C	2,541	1,374	1,167	1DANU_F140A	1,747	2,035	0,288
1CHOI_Y20D	6,197	1,597	4,6	1DANU_F147A	-0,060	0,741	0,801
1CHOI_Y20E	5,010	0,863	4,147	1DANT_G43A	0,065	0,598	0,533
1CHOI_Y20F	0,181	2,523	2,342	1DANU_G164R	-0,163	0,769	0,932
1CHOI_Y20G	3,173	1,025	2,148	1DANT_I22A	0,644	2,050	1,406
1CHOI_Y20H	1,508	0,557	0,951	1DANT_I38A	-0,126	1,941	2,067
1CHOI_Y20I	2,107	2,708	0,601	1DANT_I63A	0,000	0,559	0,559
1CHOI_Y20K	3,162	2,870	0,292	1DANU_I152A	0,179	2,969	2,79
1CHOI_Y20L	0,684	0,226	0,458	1DANH_K192A	-0,186	1,170	1,356
1CHOI_Y20M	1,225	1,522	0,297	1DANT_K15A	-0,397	2,943	3,34
1CHOI_Y20N	2,997	0,931	2,066	1DANT_K20A	2,510	2,204	0,306
1CHOI_Y20P	8,556	0,087	8,469	1DANT_K20R	1,684	2,192	0,508
1CHOI_Y20Q	3,151	0,643	2,508	1DANT_K28A	0,116	1,347	1,231
1CHOI_Y20R	2,906	2,953	0,047	1DANT_K41A	0,321	0,161	0,16
1CHOI_Y20S	2,866	0,838	2,028	1DANT_K46A	0,563	2,761	2,198
1CHOI_Y20T	3,300	1,328	1,972	1DANT_K48A	0,667	1,323	0,656
1CHOI_Y20V	3,460	1,129	2,331	1DANT_K68A	-0,070	1,333	1,403
1CHOI_Y20W	0,392	1,660	1,268	1DANU_K122A	-0,121	2,695	2,816
1CSEI_L45D	4,407	1,912	2,495	1DANU_K169A	0,116	0,478	0,362
1CSEI_L45E	2,380	1,582	0,798	1DANU_K181A	0,017	0,303	0,286
1CSEI_L45G	2,278	2,171	0,107	1DANH_L144A	0,016	0,083	0,067
1CSEI_L45I	2,979	0,112	2,867	1DANT_L59A	0,000	2,505	2,505
1CSEI_L45P	6,758	0,046	6,712	1DANT_L72A	-0,060	0,274	0,334
1CSEI_L45S	1,187	2,893	1,706	1DANU_L133A	-0,028	2,856	2,884
1DANT_D44A	1,499	2,326	0,827	1DANU_L176A	0,080	1,459	1,379
1DANT_D58A	1,987	2,954	0,967	1DANH_M164A	0,744	1,727	0,983
1DANT_D58E	1,379	1,792	0,413	1DANT_N18A	0,180	1,259	1,079
1DANT_D61A	0,242	0,097	0,145	1DANU_N199A	0,000	1,112	1,112
1DANU_D129A	-0,027	1,108	1,135	1DANU_N107A	0,000	1,783	1,783
1DANU_D145A	-0,011	2,250	2,261	1DANU_N138A	0,000	2,579	2,579
1DANT_E24A	0,657	1,928	1,271	1DANU_P92A	-0,186	1,546	1,732
1DANT_E26A	0,101	2,886	2,785	1DANT_Q37A	0,637	2,992	2,355
1DANT_E62A	0,000	2,650	2,65	1DANT_Q69A	0,000	1,322	1,322
1DANU_E208A	-0,005	1,589	1,594	1DANU_Q110A	1,304	0,635	0,669
1DANU_E105A	-0,060	2,998	3,058	1DANH_R134A	0,748	2,207	1,459
1DANU_E128A	0,086	1,221	1,135	1DANU_R196A	0,460	0,133	0,327
1DANU_E99A	-0,175	0,112	0,287	1DANU_R131A	0,000	0,041	0,041
1DANT_F50A	0,437	2,896	2,459	1DANU_R135A	0,752	2,806	2,054

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1DANU_R136A	-0,060	1,906	1,966	1DQJC_K97A	3,517	1,198	2,319
1DANU_R144A	-0,003	1,114	1,117	1DQJC_L75A	1,451	1,953	0,502
1DANT_S16A	-0,130	2,854	2,984	1DQJA_N31A	2,012	0,169	1,843
1DANT_S42A	-0,069	1,924	1,993	1DQJA_N32A	4,088	0,986	3,102
1DANT_S47A	-0,041	1,061	1,102	1DQJC_N93A	0,649	1,425	0,776
1DANU_S195A	-0,003	2,532	2,535	1DQJC_R21A	1,213	1,166	0,047
1DANU_S163A	0,023	0,476	0,453	1DQJA_S91A	1,431	1,553	0,122
1DANT_T17A	0,120	0,786	0,666	1DQJC_S100A	0,775	1,268	0,493
1DANT_T21A	-0,159	2,500	2,659	1DQJC_T89A	0,840	1,680	0,84
1DANT_T52A	0,404	1,142	0,738	1DQJB_W98A	4,928	1,596	3,332
1DANT_T60A	2,222	1,209	1,013	1DQJC_W62A	0,758	2,941	2,183
1DANU_T197A	0,110	2,619	2,509	1DQJC_W63A	1,345	0,917	0,428
1DANU_T203A	0,135	0,470	0,335	1DQJA_Y50A	2,677	1,761	0,916
1DANU_T106A	-0,060	0,490	0,55	1DQJA_Y96A	1,134	0,054	1,08
1DANU_T132A	0,000	1,042	1,042	1DQJB_Y33A	5,521	0,987	4,534
1DANU_T139A	-0,017	0,739	0,756	1DQJB_Y50A	6,883	1,775	5,108
1DANU_T167A	0,212	0,570	0,358	1DQJB_Y53A	1,179	2,914	1,735
1DANU_T172A	-0,026	0,634	0,66	1DQJC_Y20A	3,284	0,245	3,039
1DANT_V33A	-0,186	1,461	1,647	1E96A_D38N	2,200	1,482	0,718
1DANT_V36A	-0,126	2,402	2,528	1E96A_I33N	2,041	0,677	1,364
1DANT_V64A	0,000	1,627	1,627	1E96A_K132E	-0,190	1,936	2,126
1DANU_V198A	0,110	0,928	0,818	1E96A_L134R	-0,127	1,652	1,779
1DANU_V207A	0,690	2,761	2,071	1E96A_M45T	0,366	2,450	2,084
1DANU_V146A	0,199	2,020	1,821	1E96A_N26H	1,103	1,293	0,19
1DANU_V179A	0,110	1,817	1,707	1EMVA_C23A	0,921	2,116	1,195
1DANT_W25F	0,615	0,025	0,59	1EMVA_D26A	0,336	0,699	0,363
1DANT_W45A	1,933	0,673	1,26	1EMVA_D51A	5,912	0,150	5,762
1DANT_W45F	1,269	0,291	0,978	1EMVA_D60A	0,511	1,651	1,14
1DANT_W14F	0,691	2,437	1,746	1EMVA_E30A	1,415	1,806	0,391
1DANU_W158F	0,123	2,313	2,19	1EMVA_E31A	0,307	1,076	0,769
1DANT_Y51A	-0,126	1,251	1,377	1EMVA_E32A	0,220	2,965	2,745
1DANT_Y78A	0,627	2,318	1,691	1EMVA_E41A	2,082	2,594	0,512
1DANU_Y94A	0,667	1,838	1,171	1EMVA_E42A	0,657	1,075	0,418
1DANU_Y156L	0,157	2,236	2,079	1EMVA_E45A	0,213	0,593	0,38
1DANU_Y157A	0,000	2,664	2,664	1EMVB_F86A	3,876	2,738	1,138
1DANU_Y185A	-0,329	0,612	0,941	1EMVA_G49A	1,484	2,914	1,43
1DQJB_D32A	2,012	1,070	0,942	1EMVA_H46A	0,831	0,356	0,475
1DQJC_D101A	1,375	0,511	0,864	1EMVA_I53A	0,847	2,037	1,19
1DQJC_K96A	6,152	2,346	3,806	1EMVA_K35A	0,192	0,117	0,075

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1EMVB_K97A	1,958	0,474	1,484
1EMVA_L33A	3,415	2,019	1,396
1EMVA_L36A	0,906	2,673	1,767
1EMVA_L52A	0,601	1,306	0,705
1EMVA_N24A	0,139	1,513	1,374
1EMVA_N69A	0,278	0,646	0,368
1EMVB_N72A	1,164	1,691	0,527
1EMVB_N75A	2,333	0,519	1,814
1EMVA_P47A	0,437	0,520	0,083
1EMVA_P56A	1,241	1,046	0,195
1EMVB_Q92A	-0,277	0,666	0,943
1EMVB_R54A	1,665	0,195	1,47
1EMVA_S28A	0,173	1,978	1,805
1EMVA_S29A	0,955	0,304	0,651
1EMVA_S48A	0,007	1,587	1,58
1EMVA_S50A	2,186	1,021	1,165
1EMVA_S63A	0,869	0,344	0,525
1EMVB_S74A	-0,241	1,659	1,9
1EMVB_S77A	-0,233	1,262	1,495
1EMVB_S78A	-0,540	1,334	1,874
1EMVB_S84A	-0,109	2,884	2,993
1EMVA_T27A	0,728	1,416	0,688
1EMVA_T38A	0,899	1,429	0,53
1EMVA_T44A	0,304	2,895	2,591
1EMVB_T87A	0,158	1,191	1,033
1EMVA_V34A	2,577	0,270	2,307
1EMVA_V37A	1,663	0,775	0,888
1EMVA_V68A	1,855	0,067	1,788
1EMVB_V98A	1,088	1,011	0,077
1EMVA_Y54A	4,831	1,853	2,978
1EMVA_Y55A	4,632	0,230	4,402
1F47A_D4A	0,691	1,593	0,902
1F47A_D7A	1,732	1,277	0,455
1F47A_D7G	1,138	1,603	0,465
1F47A_D7S	2,063	1,183	0,88
1F47A_F11A	2,443	1,108	1,335
1F47A_I8A	2,513	1,988	0,525
1F47A_K14A	-0,043	0,620	0,663
1F47A_L12A	2,293	0,133	2,16

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1F47A_L6A	0,924	1,804	0,88
1F47A_P9A	-0,058	2,623	2,681
1F47A_Q15A	-0,046	1,704	1,75
1F47A_Y5A	0,868	0,648	0,22
1FC2C_F149A	0,014	1,130	1,116
1FC2C_F149W	0,000	0,707	0,707
1FC2C_I135W	3,135	0,143	2,992
1FC2C_I150A	3,727	1,596	2,131
1FC2C_K154A	1,378	1,549	0,171
1FC2C_L136D	1,126	2,956	1,83
1FC2C_L163W	0,000	1,135	1,135
1FC2C_N147A	0,594	2,806	2,212
1FC2C_Y133W	0,410	1,623	1,213
1FCCC_D40A	0,272	2,701	2,429
1FCCC_E42A	0,385	2,633	2,248
1FCCC_K28A	1,255	2,063	0,808
1FCCC_K31A	3,474	2,414	1,06
1FCCC_N35A	2,362	2,883	0,521
1FCCC_T25A	0,240	1,173	0,933
1FCCC_W43A	3,769	0,216	3,553
1FY8E_D194N	0,170	0,560	0,39
1FY8E_Q156K	1,170	2,151	0,981
1GCQC_A632G	1,352	0,075	1,277
1GCQC_G611V	-0,025	2,921	2,946
1GCQC_P608A	0,121	2,382	2,261
1GCQC_P609A	0,085	0,146	0,061
1GCQC_P595A	0,767	2,582	1,815
1GCQC_P657A	1,315	0,475	0,84
1GCQC_W637Y	2,145	0,630	1,515
1GL1I_L30V	4,254	1,724	2,53
1IARA_E19A	-0,320	0,731	1,051
1IARA_E19R	-0,117	1,483	1,6
1IARA_E9Q	3,109	2,222	0,887
1IARA_F82A	-0,086	0,662	0,748
1IARA_F82D	-0,580	2,642	3,222
1IARA_I11A	0,069	2,226	2,157
1IARA_I5A	1,170	0,205	0,965
1IARA_I5R	0,795	0,344	0,451
1IARA_K12E	0,139	1,108	0,969

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1IARA_K12S	-0,015	1,508	1,523
1IARA_K77A	0,154	1,215	1,061
1IARA_K77E	0,611	1,553	0,942
1IARA_K84A	0,345	1,117	0,772
1IARA_K84D	1,878	0,926	0,952
1IARA_N15A	-0,034	1,280	1,314
1IARA_N15D	-0,078	0,621	0,699
1IARA_N89A	1,557	2,265	0,708
1IARA_Q78A	0,125	0,463	0,338
1IARA_Q78E	0,245	1,519	1,274
1IARA_Q8A	-0,022	0,208	0,23
1IARA_Q8R	0,039	2,484	2,445
1IARA_R53Q	0,835	2,793	1,958
1IARA_R81A	0,479	0,512	0,033
1IARA_R81E	1,460	0,184	1,276
1IARA_R85A	0,426	2,382	1,956
1IARA_R85E	1,223	2,360	1,137
1IARA_R88A	3,751	1,655	2,096
1IARA_R88Q	2,826	1,194	1,632
1IARA_S16A	-0,183	1,005	1,188
1IARA_S16D	-0,104	1,442	1,546
1IARA_T13A	0,977	1,195	0,218
1IARA_T13D	-0,218	1,648	1,866
1IARA_T6A	-0,104	1,657	1,761
1IARA_T6D	1,391	0,338	1,053
1IARA_W91A	0,729	2,722	1,993
1IARA_W91D	1,305	0,597	0,708
1JCKB_F176A	2,132	0,817	1,315
1JCKB_K103A	0,676	2,292	1,616
1JCKB_N60A	1,641	0,012	1,629
1JCKB_T20A	1,653	2,184	0,531
1JCKB_V91A	2,230	1,883	0,347
1JCKB_Y26A	1,773	2,861	1,088
1JCKB_Y90A	2,593	2,658	0,065
1JTGB_D163A	-1,339	1,707	3,046
1JTGB_D163K	-1,981	0,562	2,543
1JTGB_D49A	1,958	0,687	1,271
1JTGA_E168A	-0,073	1,748	1,821
1JTGA_E104A	1,658	0,681	0,977

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1JTGA_E104K	4,231	1,161	3,07
1JTGA_E110A	4,057	0,760	3,297
1JTGB_F142A	2,660	0,320	2,34
1JTGB_F36A	3,198	2,319	0,879
1JTGA_G238S	-1,630	1,636	3,266
1JTGB_H148A	2,745	1,223	1,522
1JTGB_H41A	3,246	1,208	2,038
1JTGA_K234A	1,221	2,369	1,148
1JTGB_K74A	3,688	2,222	1,466
1JTGA_M129A	0,738	1,004	0,266
1JTGA_N100A	-0,455	1,351	1,806
1JTGB_N89K	-0,454	0,746	1,2
1JTGA_P107A	-0,382	2,676	3,058
1JTGA_Q99A	0,429	2,140	1,711
1JTGA_R243A	1,301	2,674	1,373
1JTGB_R160A	2,220	0,948	1,272
1JTGA_S130A	0,562	2,717	2,155
1JTGA_S235A	1,238	1,744	0,506
1JTGB_S113A	-0,168	0,567	0,735
1JTGB_S71A	0,358	0,602	0,244
1JTGB_T140K	-0,014	1,311	1,325
1JTGB_T32K	0,199	1,746	1,547
1JTGA_V216A	-0,406	1,800	2,206
1JTGA_V103A	1,909	0,763	1,146
1JTGB_V93K	-0,475	1,413	1,888
1JTGB_W112A	3,007	1,777	1,23
1JTGB_W150A	4,249	1,646	2,603
1JTGB_W162A	2,338	2,900	0,562
1JTGA_Y105A	-0,168	0,559	0,727
1JTGB_Y143A	0,382	0,098	0,284
1JTGB_Y50A	-0,406	1,811	2,217
1JTGB_Y53A	2,075	2,762	0,687
1KACA_P417S	-0,792	2,376	3,168
1KACA_S489Y	-1,251	0,594	1,845
1KTZB_D32A	1,966	1,842	0,124
1KTZB_D32N	2,444	1,428	1,016
1KTZB_D118A	1,260	1,614	0,354
1KTZB_E55A	1,661	0,784	0,877
1KTZB_E75A	1,525	2,081	0,556

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1KTZB_E119A	1,939	2,054	0,115	1PPFI_A15D	1,137	0,442	0,695
1KTZB_E119Q	2,071	1,025	1,046	1PPFI_A15E	0,105	2,152	2,047
1KTZB_F30A	3,423	1,962	1,461	1PPFI_A15F	-0,073	0,433	0,506
1KTZB_F110A	1,377	0,737	0,64	1PPFI_A15G	1,137	1,707	0,57
1KTZB_H79A	0,741	0,800	0,059	1PPFI_A15H	0,552	1,239	0,687
1KTZB_I125A	0,985	1,789	0,804	1PPFI_A15I	0,329	1,071	0,742
1KTZB_I50A	2,340	0,545	1,795	1PPFI_A15K	1,126	1,638	0,512
1KTZB_I53A	1,815	0,359	1,456	1PPFI_A15L	0,018	1,979	1,961
1KTZB_L27A	2,269	0,158	2,111	1PPFI_A15M	-0,401	1,982	2,383
1KTZB_M112A	1,317	2,714	1,397	1PPFI_A15N	0,914	0,973	0,059
1KTZB_N47A	0,731	1,094	0,363	1PPFI_A15P	1,993	1,917	0,076
1KTZA_R25A	1,480	0,329	1,151	1PPFI_A15Q	0,249	0,665	0,416
1KTZA_R25K	1,150	1,361	0,211	1PPFI_A15R	-0,264	0,576	0,84
1KTZA_R94A	2,881	2,821	0,06	1PPFI_A15S	0,756	0,883	0,127
1KTZA_R94K	2,199	1,019	1,18	1PPFI_A15T	0,962	2,688	1,726
1KTZB_S49A	0,772	2,635	1,863	1PPFI_A15V	0,200	1,792	1,592
1KTZB_S52A	0,662	1,117	0,455	1PPFI_A15W	-0,640	0,895	1,535
1KTZB_S52L	4,479	0,680	3,799	1PPFI_A15Y	-0,185	2,099	2,284
1KTZB_T51A	1,958	0,048	1,91	1PPFI_E19A	1,202	1,808	0,606
1KTZA_V92I	0,243	1,200	0,957	1PPFI_E19C	1,401	2,745	1,344
1KTZB_V62A	1,093	2,855	1,762	1PPFI_E19D	0,577	2,299	1,722
1KTZB_V77A	0,861	1,674	0,813	1PPFI_E19F	1,363	0,263	1,1
1MAHA_D74N	1,733	2,808	1,075	1PPFI_E19G	2,118	0,417	1,701
1MAHA_F295L	1,148	0,015	1,133	1PPFI_E19H	0,690	1,297	0,607
1MAHA_F297I	1,900	0,652	1,248	1PPFI_E19I	0,723	2,231	1,508
1MAHA_F297Y	0,730	0,215	0,515	1PPFI_E19K	2,118	1,267	0,851
1MAHA_F338G	0,730	2,120	1,39	1PPFI_E19L	1,070	2,641	1,571
1MAHA_W286R	8,122	0,954	7,168	1PPFI_E19M	1,147	0,403	0,744
1MAHA_Y124Q	2,770	1,409	1,361	1PPFI_E19N	1,211	0,095	1,116
1MAHA_Y337A	0,356	1,185	0,829	1PPFI_E19P	3,207	2,264	0,943
1MAHA_Y72N	4,811	1,699	3,112	1PPFI_E19Q	0,660	2,033	1,373
1N8OE_M84K	1,458	1,938	0,48	1PPFI_E19R	1,468	0,375	1,093
1N8OE_M84L	-0,062	1,594	1,656	1PPFI_E19S	1,844	1,777	0,067
1N8OE_M84R	1,348	2,154	0,806	1PPFI_E19T	1,503	1,760	0,257
1NCAH_D97K	0,714	2,990	2,276	1PPFI_E19V	1,158	0,470	0,688
1NCAH_E96D	0,410	2,488	2,078	1PPFI_E19W	1,528	0,071	1,457
1NCAH_N98Q	0,542	0,470	0,072	1PPFI_E19Y	1,247	1,943	0,696
1NCAH_N31Q	0,000	0,886	0,886	1PPFI_G32A	0,257	0,504	0,247
1PPFI_A15C	-0,673	0,020	0,693	1PPFI_G32C	1,178	2,692	1,514

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1PPFI_G32D	2,494	2,945	0,451
1PPFI_G32E	1,456	0,649	0,807
1PPFI_G32F	1,480	0,409	1,071
1PPFI_G32H	1,503	0,101	1,402
1PPFI_G32I	4,367	2,278	2,089
1PPFI_G32K	3,303	2,087	1,216
1PPFI_G32L	3,126	0,601	2,525
1PPFI_G32M	2,193	2,742	0,549
1PPFI_G32N	1,612	2,910	1,298
1PPFI_G32P	0,264	0,321	0,057
1PPFI_G32Q	2,797	1,874	0,923
1PPFI_G32R	4,844	2,410	2,434
1PPFI_G32S	0,914	0,427	0,487
1PPFI_G32T	2,809	0,907	1,902
1PPFI_G32V	2,545	0,253	2,292
1PPFI_G32W	1,642	1,542	0,1
1PPFI_G32Y	1,202	0,954	0,248
1PPFI_K13A	0,756	0,143	0,613
1PPFI_K13C	0,962	0,853	0,109
1PPFI_K13D	0,631	0,836	0,205
1PPFI_K13E	1,132	1,619	0,487
1PPFI_K13F	0,962	2,590	1,628
1PPFI_K13G	1,247	2,094	0,847
1PPFI_K13H	1,013	2,697	1,684
1PPFI_K13I	0,631	1,205	0,574
1PPFI_K13L	0,401	0,934	0,533
1PPFI_K13M	0,208	1,056	0,848
1PPFI_K13N	0,660	2,477	1,817
1PPFI_K13P	1,265	0,216	1,049
1PPFI_K13Q	0,249	2,648	2,399
1PPFI_K13R	-0,640	0,593	1,233
1PPFI_K13S	0,460	1,023	0,563
1PPFI_K13T	0,166	0,335	0,169
1PPFI_K13V	0,871	1,605	0,734
1PPFI_K13W	0,830	1,020	0,19
1PPFI_K13Y	0,505	0,286	0,219
1PPFI_L18A	1,016	1,369	0,353
1PPFI_L18C	-0,089	2,986	3,075
1PPFI_L18D	7,605	1,825	5,78

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1PPFI_L18E	7,056	0,167	6,889
1PPFI_L18F	5,164	1,358	3,806
1PPFI_L18G	3,329	0,291	3,038
1PPFI_L18H	6,464	0,381	6,083
1PPFI_L18I	-0,732	1,031	1,763
1PPFI_L18K	5,687	1,014	4,673
1PPFI_L18M	1,240	1,975	0,735
1PPFI_L18N	5,220	0,833	4,387
1PPFI_L18P	6,140	1,241	4,899
1PPFI_L18Q	3,344	0,706	2,638
1PPFI_L18R	7,177	2,808	4,369
1PPFI_L18S	3,089	2,921	0,168
1PPFI_L18T	0,914	0,182	0,732
1PPFI_L18V	-0,492	1,148	1,64
1PPFI_L18W	7,536	2,192	5,344
1PPFI_L18Y	6,628	1,518	5,11
1PPFI_N36A	-1,644	1,014	2,658
1PPFI_N36C	0,604	0,514	0,09
1PPFI_N36D	-3,075	2,530	5,605
1PPFI_N36E	-1,017	2,105	3,122
1PPFI_N36F	1,844	2,937	1,093
1PPFI_N36G	-0,571	2,371	2,942
1PPFI_N36H	-0,532	0,044	0,576
1PPFI_N36I	0,871	2,107	1,236
1PPFI_N36K	2,619	1,401	1,218
1PPFI_N36L	2,644	0,684	1,96
1PPFI_N36M	1,093	2,219	1,126
1PPFI_N36P	-2,956	0,040	2,996
1PPFI_N36Q	0,382	2,546	2,164
1PPFI_N36R	1,868	0,609	1,259
1PPFI_N36S	-1,345	1,409	2,754
1PPFI_N36T	0,552	2,217	1,665
1PPFI_N36V	0,345	0,856	0,511
1PPFI_N36W	1,726	0,630	1,096
1PPFI_N36Y	1,675	0,582	1,093
1PPFI_P14A	-0,124	1,019	1,143
1PPFI_P14C	-2,003	0,352	2,355
1PPFI_P14D	-0,492	1,700	2,192
1PPFI_P14E	-1,427	1,445	2,872

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1PPFI_P14F	-1,855	2,125	3,98
1PPFI_P14G	0,093	0,291	0,198
1PPFI_P14H	-1,764	1,080	2,844
1PPFI_P14I	-1,650	1,527	3,177
1PPFI_P14K	-0,448	1,394	1,842
1PPFI_P14L	-2,920	2,388	5,308
1PPFI_P14M	-1,573	1,761	3,334
1PPFI_P14N	-0,640	1,639	2,279
1PPFI_P14Q	-0,810	2,912	3,722
1PPFI_P14R	-0,137	0,642	0,779
1PPFI_P14S	-0,571	0,473	1,044
1PPFI_P14T	-0,264	0,440	0,704
1PPFI_P14V	-1,523	0,780	2,303
1PPFI_P14W	-2,003	2,583	4,586
1PPFI_P14Y	-0,962	1,572	2,534
1PPFI_R21A	0,208	0,113	0,095
1PPFI_R21C	-0,081	1,849	1,93
1PPFI_R21D	0,208	2,905	2,697
1PPFI_R21E	0,460	2,978	2,518
1PPFI_R21F	-0,980	0,176	1,156
1PPFI_R21G	0,577	0,947	0,37
1PPFI_R21H	-0,492	1,205	1,697
1PPFI_R21I	-0,810	0,187	0,997
1PPFI_R21K	0,604	0,318	0,286
1PPFI_R21L	-0,858	1,027	1,885
1PPFI_R21M	-0,759	1,184	1,943
1PPFI_R21N	0,329	2,837	2,508
1PPFI_R21P	6,698	1,632	5,066
1PPFI_R21Q	-0,018	2,651	2,669
1PPFI_R21S	0,419	2,360	1,941
1PPFI_R21T	-0,018	0,846	0,864
1PPFI_R21V	-0,349	0,149	0,498
1PPFI_R21W	-0,532	1,208	1,74
1PPFI_R21Y	0,220	2,592	2,372
1PPFI_T17A	3,186	0,337	2,849
1PPFI_T17C	2,085	2,908	0,823
1PPFI_T17D	4,918	1,528	3,39
1PPFI_T17E	3,055	1,915	1,14
1PPFI_T17F	1,708	1,992	0,284

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1PPFI_T17G	3,596	1,216	2,38
1PPFI_T17H	1,692	0,101	1,591
1PPFI_T17I	1,492	2,861	1,369
1PPFI_T17K	3,385	1,999	1,386
1PPFI_T17L	2,277	1,415	0,862
1PPFI_T17M	2,500	1,075	1,425
1PPFI_T17N	2,696	1,131	1,565
1PPFI_T17P	3,071	2,431	0,64
1PPFI_T17Q	1,967	2,763	0,796
1PPFI_T17R	3,448	0,522	2,926
1PPFI_T17S	1,529	0,320	1,209
1PPFI_T17V	1,511	0,034	1,477
1PPFI_T17W	2,193	2,211	0,018
1PPFI_T17Y	2,478	1,954	0,524
1PPFI_Y20A	3,207	0,134	3,073
1PPFI_Y20C	3,448	0,810	2,638
1PPFI_Y20D	6,412	0,647	5,765
1PPFI_Y20E	6,412	0,804	5,608
1PPFI_Y20F	0,482	2,082	1,6
1PPFI_Y20G	4,181	1,996	2,185
1PPFI_Y20H	2,413	0,733	1,68
1PPFI_Y20I	3,927	0,679	3,248
1PPFI_Y20K	4,489	1,196	3,293
1PPFI_Y20L	1,373	0,945	0,428
1PPFI_Y20M	2,809	1,134	1,675
1PPFI_Y20N	3,640	2,834	0,806
1PPFI_Y20P	5,361	1,770	3,591
1PPFI_Y20Q	4,692	0,349	4,343
1PPFI_Y20R	4,367	2,432	1,935
1PPFI_Y20S	3,640	2,115	1,525
1PPFI_Y20T	5,050	0,280	4,77
1PPFI_Y20V	4,181	1,054	3,127
1PPFI_Y20W	0,235	1,429	1,194
1R0RI_A15C	-0,483	0,982	1,465
1R0RI_A15D	5,223	0,624	4,599
1R0RI_A15E	4,573	0,175	4,398
1R0RI_A15F	-2,240	2,004	4,244
1R0RI_A15G	1,110	0,495	0,615
1R0RI_A15H	1,773	2,464	0,691

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1R0RI_A15I	-1,798	1,834	3,632
1R0RI_A15K	3,070	1,779	1,291
1R0RI_A15L	-1,753	0,394	2,147
1R0RI_A15M	-1,738	2,633	4,371
1R0RI_A15N	2,093	2,982	0,889
1R0RI_A15P	3,342	1,010	2,332
1R0RI_A15Q	1,809	2,105	0,296
1R0RI_A15R	5,524	1,495	4,029
1R0RI_A15S	1,740	1,162	0,578
1R0RI_A15T	1,267	1,322	0,055
1R0RI_A15V	-0,969	0,124	1,093
1R0RI_A15W	0,207	2,654	2,447
1R0RI_A15Y	-0,838	0,901	1,739
1R0RI_E19A	2,087	2,540	0,453
1R0RI_E19C	2,369	0,997	1,372
1R0RI_E19D	0,258	0,365	0,107
1R0RI_E19F	3,614	1,836	1,778
1R0RI_E19G	2,709	2,082	0,627
1R0RI_E19H	1,707	1,905	0,198
1R0RI_E19I	0,736	0,280	0,456
1R0RI_E19K	2,709	1,682	1,027
1R0RI_E19L	0,569	1,572	1,003
1R0RI_E19M	1,569	2,812	1,243
1R0RI_E19N	2,116	0,345	1,771
1R0RI_E19P	3,760	2,291	1,469
1R0RI_E19Q	1,281	1,418	0,137
1R0RI_E19R	2,093	0,218	1,875
1R0RI_E19S	3,172	2,837	0,335
1R0RI_E19T	4,248	1,531	2,717
1R0RI_E19V	0,095	2,144	2,049
1R0RI_E19W	1,847	0,126	1,721
1R0RI_E19Y	1,888	0,200	1,688
1R0RI_G32A	1,314	0,623	0,691
1R0RI_G32C	1,418	2,513	1,095
1R0RI_G32D	2,885	1,695	1,19
1R0RI_G32E	1,979	0,938	1,041
1R0RI_G32F	0,115	2,605	2,49
1R0RI_G32H	2,190	1,211	0,979
1R0RI_G32I	1,979	1,240	0,739

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1R0RI_G32K	2,932	2,565	0,367
1R0RI_G32L	2,408	0,105	2,303
1R0RI_G32M	2,154	1,831	0,323
1R0RI_G32N	1,281	2,840	1,559
1R0RI_G32P	1,079	2,707	1,628
1R0RI_G32Q	2,031	2,016	0,015
1R0RI_G32R	3,467	1,959	1,508
1R0RI_G32S	0,911	0,747	0,164
1R0RI_G32T	1,677	0,858	0,819
1R0RI_G32V	1,956	2,050	0,094
1R0RI_G32W	1,400	1,674	0,274
1R0RI_G32Y	0,485	2,222	1,737
1R0RI_K13A	-0,609	0,088	0,697
1R0RI_K13C	-0,615	2,773	3,388
1R0RI_K13D	-0,593	1,601	2,194
1R0RI_K13E	-0,110	2,685	2,795
1R0RI_K13F	-0,794	2,621	3,415
1R0RI_K13G	-0,838	2,053	2,891
1R0RI_K13H	-0,593	2,402	2,995
1R0RI_K13I	1,123	2,850	1,727
1R0RI_K13L	-0,328	1,042	1,37
1R0RI_K13M	-0,561	2,663	3,224
1R0RI_K13N	-0,390	1,186	1,576
1R0RI_K13P	1,346	0,943	0,403
1R0RI_K13Q	-0,530	1,156	1,686
1R0RI_K13R	0,136	2,952	2,816
1R0RI_K13S	-0,307	2,290	2,597
1R0RI_K13T	0,889	2,597	1,708
1R0RI_K13V	0,410	0,114	0,296
1R0RI_K13W	-1,104	1,778	2,882
1R0RI_K13Y	-0,695	2,549	3,244
1R0RI_L18A	0,419	1,409	0,99
1R0RI_L18C	-1,345	2,399	3,744
1R0RI_L18D	4,534	1,770	2,764
1R0RI_L18E	2,148	1,651	0,497
1R0RI_L18F	0,616	2,947	2,331
1R0RI_L18G	2,350	0,780	1,57
1R0RI_L18H	0,569	1,060	0,491
1R0RI_L18I	3,250	2,961	0,289

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1R0RI_L18K	3,091	2,624	0,467
1R0RI_L18M	-0,516	1,238	1,754
1R0RI_L18N	1,499	1,319	0,18
1R0RI_L18P	7,679	2,879	4,8
1R0RI_L18Q	0,592	1,436	0,844
1R0RI_L18R	2,932	1,547	1,385
1R0RI_L18S	1,569	0,425	1,144
1R0RI_L18T	0,160	0,485	0,325
1R0RI_L18V	2,154	1,151	1,003
1R0RI_L18W	1,522	1,299	0,223
1R0RI_L18Y	0,446	0,041	0,405
1R0RI_N28S	-0,066	2,309	2,375
1R0RI_N36A	-0,033	2,422	2,455
1R0RI_N36C	0,569	2,185	1,616
1R0RI_N36D	0,569	0,657	0,088
1R0RI_N36E	-0,205	2,729	2,934
1R0RI_N36F	-0,180	2,677	2,857
1R0RI_N36G	0,285	2,197	1,912
1R0RI_N36H	0,285	2,952	2,667
1R0RI_N36I	0,314	2,169	1,855
1R0RI_N36K	1,146	1,991	0,845
1R0RI_N36L	1,545	0,448	1,097
1R0RI_N36M	0,314	2,266	1,952
1R0RI_N36P	1,478	0,984	0,494
1R0RI_N36Q	-0,274	1,125	1,399
1R0RI_N36R	1,457	2,674	1,217
1R0RI_N36S	0,485	0,994	0,509
1R0RI_N36T	0,569	2,946	2,377
1R0RI_N36V	0,668	2,749	2,081
1R0RI_N36W	0,056	1,906	1,85
1R0RI_N36Y	0,182	1,284	1,102
1R0RI_P14A	-0,638	0,702	1,34
1R0RI_P14C	-0,986	2,658	3,644
1R0RI_P14D	-2,518	2,184	4,702
1R0RI_P14E	-2,240	2,947	5,187
1R0RI_P14F	1,380	2,181	0,801
1R0RI_P14G	-0,345	2,066	2,411
1R0RI_P14H	-0,216	0,787	1,003
1R0RI_P14I	1,740	0,738	1,002

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1R0RI_P14K	0,344	1,329	0,985
1R0RI_P14L	1,183	1,432	0,249
1R0RI_P14M	-0,033	0,170	0,203
1R0RI_P14N	-0,916	2,557	3,473
1R0RI_P14Q	-1,413	0,273	1,686
1R0RI_P14R	1,101	2,696	1,595
1R0RI_P14S	0,136	0,659	0,523
1R0RI_P14T	1,196	1,208	0,012
1R0RI_P14V	1,330	1,063	0,267
1R0RI_P14W	0,485	1,690	1,205
1R0RI_P14Y	0,857	2,263	1,406
1R0RI_R21A	-0,096	0,242	0,338
1R0RI_R21C	0,115	0,423	0,308
1R0RI_R21D	-0,180	1,389	1,569
1R0RI_R21E	0,344	2,677	2,333
1R0RI_R21F	0,446	0,219	0,227
1R0RI_R21G	2,087	2,063	0,024
1R0RI_R21H	0,525	0,657	0,132
1R0RI_R21I	0,207	0,097	0,11
1R0RI_R21K	-0,096	1,515	1,611
1R0RI_R21L	0,232	1,283	1,051
1R0RI_R21M	0,446	1,870	1,424
1R0RI_R21N	-0,050	2,502	2,552
1R0RI_R21P	7,382	0,898	6,484
1R0RI_R21Q	0,182	2,984	2,802
1R0RI_R21S	-0,016	0,226	0,242
1R0RI_R21T	0,344	1,178	0,834
1R0RI_R21V	0,115	2,213	2,098
1R0RI_R21W	0,944	1,530	0,586
1R0RI_R21Y	0,822	1,013	0,191
1R0RI_T17A	0,979	0,472	0,507
1R0RI_T17C	0,812	2,322	1,51
1R0RI_T17D	2,885	1,194	1,691
1R0RI_T17E	3,015	2,006	1,009
1R0RI_T17F	-0,066	0,445	0,511
1R0RI_T17G	3,135	2,210	0,925
1R0RI_T17H	0,525	0,714	0,189
1R0RI_T17I	0,258	2,941	2,683
1R0RI_T17K	1,932	0,560	1,372

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1R0RI_T17L	0,668	2,980	2,312
1R0RI_T17M	1,330	1,220	0,11
1R0RI_T17N	0,018	0,158	0,14
1R0RI_T17P	4,944	0,132	4,812
1R0RI_T17Q	2,219	0,186	2,033
1R0RI_T17R	3,342	0,533	2,809
1R0RI_T17S	0,858	2,108	1,25
1R0RI_T17V	1,007	2,941	1,934
1R0RI_T17W	2,932	2,382	0,55
1R0RI_T17Y	-0,878	0,085	0,963
1R0RI_Y20A	5,468	2,279	3,189
1R0RI_Y20C	3,250	2,141	1,109
1R0RI_Y20D	5,616	0,868	4,748
1R0RI_Y20E	4,613	0,919	3,694
1R0RI_Y20F	0,616	1,988	1,372
1R0RI_Y20G	6,395	1,185	5,21
1R0RI_Y20H	3,253	2,960	0,293
1R0RI_Y20I	2,861	2,244	0,617
1R0RI_Y20K	5,506	2,340	3,166
1R0RI_Y20L	2,572	1,970	0,602
1R0RI_Y20M	3,135	2,827	0,308
1R0RI_Y20N	5,610	2,228	3,382
1R0RI_Y20P	6,647	2,659	3,988
1R0RI_Y20Q	4,498	0,611	3,887
1R0RI_Y20R	4,465	1,080	3,385
1R0RI_Y20S	5,157	0,564	4,593
1R0RI_Y20T	5,468	2,579	2,889
1R0RI_Y20V	3,506	2,204	1,302
1R0RI_Y20W	-0,364	0,282	0,646
1S1QA_D46A	0,964	0,798	0,166
1S1QA_F44A	0,199	0,145	0,054
1S1QA_F88A	0,774	1,330	0,556
1S1QA_N45A	1,231	0,217	1,014
1S1QA_V43A	0,670	0,095	0,575
1S1QA_W75A	0,280	2,822	2,542
1SBBB_L20T	-0,091	1,824	1,915
1SBBB_V26Y	-1,454	0,656	2,11
1SBBB_Y91V	0,079	0,806	0,727
1TM1I_E60A	2,986	2,064	0,922

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1TM1I_E60S	2,745	1,902	0,843
1TM1I_M59A	1,027	0,316	0,711
1TM1I_M59F	1,027	1,874	0,847
1TM1I_M59G	2,230	2,425	0,195
1TM1I_M59K	1,092	0,500	0,592
1TM1I_M59Y	0,056	1,225	1,169
1TM1I_R62A	1,255	1,626	0,371
1TM1I_R65A	3,414	1,453	1,961
1TM1I_R67A	3,008	2,387	0,621
1TM1I_R67C	3,232	1,579	1,653
1TM1I_T58A	2,648	0,732	1,916
1TM1I_T58D	2,097	1,924	0,173
1TM1I_T58P	3,752	1,424	2,328
1TM1I_V70A	0,025	1,348	1,323
1TM1I_Y61A	2,577	2,468	0,109
1TM1I_Y61G	4,676	2,298	2,378
1UUZA_C64A	0,650	1,083	0,433
1UUZA_H62A	1,773	1,522	0,251
1UUZA_H62D	3,375	1,363	2,012
1UUZA_H62N	1,518	2,248	0,73
1UUZA_H62Q	1,518	1,151	0,367
1XD3B_D39A	-0,410	2,011	2,421
1XD3B_D52A	-0,240	0,602	0,842
1XD3B_D58A	-0,240	2,975	3,215
1XD3B_E24A	-0,240	1,072	1,312
1XD3B_E51A	1,533	2,434	0,901
1XD3B_H68N	0,000	2,954	2,954
1XD3B_I44A	0,266	1,468	1,202
1XD3B_K11R	1,401	1,478	0,077
1XD3B_K27A	-0,062	2,986	3,048
1XD3B_K27R	0,266	1,498	1,232
1XD3B_K33R	0,000	1,531	1,531
1XD3B_K6A	1,343	0,159	1,184
1XD3B_K6R	0,302	2,204	1,902
1XD3B_L8A	2,674	1,542	1,132
1XD3B_R42L	-0,861	1,099	1,96
1XD3B_R54L	0,868	0,870	0,002
1XD3B_R72L	1,300	1,054	0,246
1XD3B_R74L	2,375	2,657	0,282

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1Z7XW_D435A	3,658	1,123	2,535	2I9BE_K139A	0,674	1,942	1,268
1Z7XW_E206A	1,018	0,644	0,374	2I9BE_R137A	-0,287	1,878	2,165
1Z7XW_E287A	1,320	2,850	1,53	2I9BE_R142A	0,363	0,566	0,203
1Z7XW_E344A	1,560	0,318	1,242	2I9BE_R145A	0,420	0,197	0,223
1Z7XW_E401A	1,305	2,040	0,735	2J0TD_C70S	4,304	2,284	2,02
1Z7XW_I459A	0,337	0,247	0,09	2J0TD_M66A	1,641	1,832	0,191
1Z7XW_K320A	1,320	1,114	0,206	2J0TD_S68A	2,104	2,307	0,203
1Z7XW_R457A	0,847	1,829	0,982	2J0TD_S68E	2,183	0,038	2,145
1Z7XW_S289A	0,813	2,805	1,992	2J0TD_S68R	2,833	2,270	0,563
1Z7XW_W261A	1,334	2,537	1,203	2J0TD_S68Y	2,975	2,234	0,741
1Z7XW_W263A	2,210	1,272	0,938	2J0TD_T2A	4,284	1,361	2,923
1Z7XW_W318A	0,992	1,747	0,755	2J0TD_T2L	2,679	0,101	2,578
1Z7XW_W375A	1,668	1,921	0,253	2J0TD_T2R	5,040	2,425	2,615
1Z7XW_Y434A	5,949	1,362	4,587	2J0TD_T2S	1,593	2,819	1,226
1Z7XW_Y434F	0,120	1,049	0,929	2J0TD_V4A	0,000	0,821	0,821
1Z7XW_Y437A	2,621	1,157	1,464	2J0TD_V4I	1,641	1,647	0,006
1Z7XW_Y437F	2,164	2,638	0,474	2J0TD_V4K	1,822	2,771	0,949
2A9KB_G99D	2,315	0,721	1,594	2J0TD_V4S	0,981	2,916	1,935
2AJFE_F360S	-0,082	1,690	1,772	2J0TD_V69I	0,790	0,265	0,525
2AJFE_K344R	-0,007	0,288	0,295	2J0TD_V69T	-0,076	1,576	1,652
2AJFE_N479K	2,010	2,367	0,357	2JELP_A82S	0,000	1,087	1,087
2AJFE_T487S	1,822	1,974	0,152	2JELP_D69E	0,952	0,709	0,243
2BTFP_F59A	1,594	2,768	1,174	2JELP_E5D	0,410	0,280	0,13
2BTFP_G120F	1,719	1,916	0,197	2JELP_E5Q	0,713	2,276	1,563
2BTFP_K125A	0,460	1,278	0,818	2JELP_E66K	4,088	1,540	2,548
2BTFP_V60E	0,978	0,643	0,335	2JELP_E68A	0,410	0,873	0,463
2C0LA_N382A	0,687	2,381	1,694	2JELP_E70A	2,725	2,742	0,017
2C0LA_Q586R	1,485	0,975	0,51	2JELP_E70K	4,088	2,093	1,995
2C0LA_S589Y	0,273	0,730	0,457	2JELP_E75R	2,725	2,237	0,488
2GOXB_N138A	1,450	0,728	0,722	2JELP_E83A	0,000	1,908	1,908
2GOXB_R131A	1,881	1,448	0,433	2JELP_E85A	0,000	2,830	2,83
2HLEA_K149Q	-0,410	2,057	2,467	2JELP_E85D	0,000	2,424	2,424
2HLEA_L95R	2,285	2,942	0,657	2JELP_E85K	0,000	0,632	0,632
2HRKA_K159A	0,952	2,538	1,586	2JELP_E85Q	0,000	1,887	1,887
2HRKA_T127V	0,720	0,866	0,146	2JELP_F2W	2,617	1,538	1,079
2HRKB_T57V	0,624	2,713	2,089	2JELP_F2Y	0,000	2,031	2,031
2I26N_A30V	0,030	1,968	1,938	2JELP_H76A	-0,410	2,541	2,951
2I26N_S61R	-0,626	1,703	2,329	2JELP_H76D	-0,650	0,611	1,261
2I9BE_H143A	0,661	2,610	1,949	2JELP_K24E	0,000	1,433	1,433

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
2JELP_K27E	0,000	2,330	2,33
2JELP_K72E	0,410	1,354	0,944
2JELP_K72R	0,000	2,778	2,778
2JELP_K79E	0,410	0,827	0,417
2JELP_N12D	0,000	1,802	1,802
2JELP_N38T	0,000	0,530	0,53
2JELP_P11E	0,000	0,241	0,241
2JELP_Q3K	4,088	2,617	1,471
2JELP_Q4K	1,363	0,364	0,999
2JELP_Q57E	-0,410	2,966	3,376
2JELP_Q71E	2,725	1,739	0,986
2JELP_R17G	0,000	2,796	2,796
2JELP_R17K	0,000	2,765	2,765
2JELP_S41C	1,495	2,435	0,94
2JELP_S43C	0,000	0,882	0,882
2JELP_S46C	0,000	0,103	0,103
2JELP_S64T	4,088	0,870	3,218
2JELP_T34Q	0,000	1,041	1,041
2JELP_T36Q	0,410	2,597	2,187
2JELP_T62A	0,000	0,860	0,86
2JELP_T62N	0,000	2,511	2,511
2JELP_T7N	0,410	0,973	0,563
2JELP_T7S	0,000	0,331	0,331
2JELP_V6F	0,000	1,738	1,738
2SICI_M73A	0,218	1,696	1,478
2SICI_M73D	0,751	0,267	0,484
2SICI_M73E	0,795	2,244	1,449
2SICI_M73G	0,145	1,423	1,278
2SICI_M73H	0,218	0,380	0,162
2SICI_M73I	1,603	0,631	0,972
2SICI_M73K	0,000	2,017	2,017
2SICI_M73L	-0,240	2,190	2,43
2SICI_M73R	0,000	1,901	1,901
2SICI_M73V	0,713	2,933	2,22
2VIRC_S157L	3,678	2,961	0,717
2VIRC_T131I	4,908	0,007	4,901
2VLJE_D32A	1,571	0,152	1,419
2VLJE_D56A	0,132	0,739	0,607
2VLJE_I53L	1,417	0,238	1,179

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
2VLJE_I53V	0,201	1,972	1,771
2VLJE_N55A	1,128	0,147	0,981
2VLJE_N55D	0,495	0,819	0,324
2VLJD_Q34A	0,975	0,655	0,32
2VLJE_Q58A	0,495	0,818	0,323
2VLJE_Q58E	0,000	2,123	2,123
2VLJD_S31A	0,627	2,164	1,537
2VLJD_S32A	1,037	1,450	0,413
2VLJE_S99A	-0,035	0,759	0,794
2VLJE_Y101A	0,232	2,443	2,211
2VLJE_Y101F	0,495	0,939	0,444
2WPTA_D33A	-0,132	0,364	0,496
2WPTA_D33L	-3,400	2,006	5,406
2WPTA_E30A	1,732	2,935	1,203
2WPTA_E39H	0,038	2,660	2,622
2WPTA_E41A	4,498	1,495	3,003
2WPTB_F86A	1,057	2,495	1,438
2WPTB_K97A	0,649	1,988	1,339
2WPTA_N34A	-0,379	2,454	2,833
2WPTA_N34V	-0,896	0,307	1,203
2WPTB_N72A	0,701	0,174	0,527
2WPTB_N75A	1,250	2,948	1,698
2WPTA_P56A	2,924	2,217	0,707
2WPTB_Q92A	0,383	2,244	1,861
2WPTA_R38A	-1,110	1,568	2,678
2WPTA_R38T	-1,037	1,106	2,143
2WPTA_R42A	-0,240	0,826	1,066
2WPTA_R42E	-0,372	0,814	1,186
2WPTB_R54A	0,871	1,593	0,722
2WPTA_S50A	2,423	2,521	0,098
2WPTB_S74A	-0,134	1,828	1,962
2WPTB_S77A	-0,456	1,574	2,03
2WPTB_S78A	-0,095	2,388	2,483
2WPTB_S84A	-0,067	1,216	1,283
2WPTB_T87A	0,377	1,915	1,538
2WPTA_V37A	3,805	2,930	0,875
2WPTB_V98A	0,264	2,904	2,64
3BK3C_A36R	1,381	2,732	1,351
3BK3C_I18A	0,486	1,944	1,458

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3BK3C_I18R	0,423	1,528	1,105	3HFMH_D32N	0,170	1,140	0,97
3BK3C_I21A	1,306	1,806	0,5	3HFMY_D101A	1,227	2,315	1,088
3BK3C_I21R	2,119	1,445	0,674	3HFMY_D101E	2,132	2,156	0,024
3BK3C_I27A	1,260	1,807	0,547	3HFMY_D101F	2,326	0,835	1,491
3BK3C_I27R	1,174	1,699	0,525	3HFMY_D101G	0,359	0,708	0,349
3BK3C_I2A	1,037	0,075	0,962	3HFMY_D101K	2,123	1,034	1,089
3BK3C_I2R	0,650	1,277	0,627	3HFMY_D101N	1,490	0,046	1,444
3BK3C_L1A	0,000	0,160	0,16	3HFMY_D101Q	2,083	0,127	1,956
3BK3C_L1R	0,026	2,968	2,942	3HFMY_D101R	2,279	0,497	1,782
3BK3C_T3P	0,486	2,361	1,875	3HFMY_D101S	1,868	2,105	0,237
3BK3C_T5P	1,741	2,902	1,161	3HFMY_G102V	0,395	0,033	0,362
3BN9B_D217A	0,566	1,425	0,859	3HFMY_H15A	-0,445	2,852	3,297
3BN9B_D60AA	0,422	1,422	1	3HFMY_I98A	0,000	2,160	2,16
3BN9B_D60BA	0,311	2,833	2,522	3HFMY_K96A	6,983	2,246	4,737
3BN9B_E169A	0,373	0,898	0,525	3HFMY_K96M	6,772	1,748	5,024
3BN9B_F60EA	-0,045	1,993	2,038	3HFMY_K96R	5,349	2,004	3,345
3BN9B_F94A	0,639	1,269	0,63	3HFMY_K97A	5,858	1,776	4,082
3BN9B_H143A	0,085	0,369	0,284	3HFMY_K97D	6,764	2,866	3,898
3BN9B_I41A	0,000	1,036	1,036	3HFMY_K97E	3,602	0,002	3,6
3BN9B_I60A	0,835	1,072	0,237	3HFMY_K97G	6,426	0,413	6,013
3BN9B_K224A	0,784	2,255	1,471	3HFMY_K97M	0,946	2,059	1,113
3BN9B_L153A	0,336	2,058	1,722	3HFMY_K97R	3,051	0,056	2,995
3BN9B_N95A	0,773	0,524	0,249	3HFMY_L75A	0,704	0,101	0,603
3BN9B_Q145A	0,133	2,446	2,313	3HFML_N31A	5,211	0,339	4,872
3BN9B_Q174A	-0,035	1,660	1,695	3HFML_N31D	1,343	1,390	0,047
3BN9B_Q175A	2,507	0,960	1,547	3HFML_N31E	5,704	2,933	2,771
3BN9B_Q221AA	0,705	2,819	2,114	3HFML_N32A	5,104	1,496	3,608
3BN9B_Q38A	-0,415	2,215	2,63	3HFMY_N19D	0,410	1,679	1,269
3BN9B_R222A	-0,094	2,619	2,713	3HFMY_N19K	0,240	0,910	0,67
3BN9B_R60CA	-0,045	0,449	0,494	3HFMY_N19Q	-0,042	2,513	2,555
3BN9B_R60FA	-0,072	0,388	0,46	3HFMY_N93A	0,211	0,832	0,621
3BN9B_R87A	-0,159	0,594	0,753	3HFML_Q53A	0,952	2,623	1,671
3BN9B_T150A	0,291	1,806	1,515	3HFMY_R21A	0,903	1,617	0,714
3BN9B_T98A	1,131	1,246	0,115	3HFMY_R21E	2,452	0,216	2,236
3BN9B_Y146A	1,084	0,815	0,269	3HFMY_R21G	2,419	2,231	0,188
3BN9B_Y60GA	0,019	0,336	0,317	3HFMY_R21H	2,168	1,505	0,663
3BP8C_A63F	0,615	2,235	1,62	3HFMY_R21K	1,758	0,832	0,926
3BP8A_F136A	0,708	1,167	0,459	3HFMY_R21M	2,049	2,646	0,597
3HFMH_D32A	1,897	2,128	0,231	3HFMY_R21N	2,326	1,732	0,594

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3HFMY_R21Q	2,396	0,722	1,674	3NPSA_H143A	1,874	0,798	1,076
3HFMY_R21W	2,132	1,417	0,715	3NPSA_I41A	0,641	0,598	0,043
3HFMY_R73A	-0,331	1,917	2,248	3NPSA_I60A	-0,332	0,595	0,927
3HFMH_S31A	0,170	2,334	2,164	3NPSA_K224A	-0,104	1,181	1,285
3HFMY_S100A	0,267	2,920	2,653	3NPSA_L153A	0,296	1,121	0,825
3HFMY_T89A	0,000	1,596	1,596	3NPSA_N95A	0,253	2,060	1,807
3HFMH_W98A	5,507	2,220	3,287	3NPSA_Q145A	0,296	0,938	0,642
3HFMH_W98F	3,243	0,313	2,93	3NPSA_Q174A	-0,059	1,509	1,568
3HFMY_W63A	0,319	0,905	0,586	3NPSA_Q175A	0,740	1,733	0,993
3HFMH_Y33A	6,031	0,584	5,447	3NPSA_Q221AA	-0,041	1,137	1,178
3HFMH_Y33F	1,053	0,208	0,845	3NPSA_Q38A	0,026	0,488	0,462
3HFMH_Y33L	1,973	2,288	0,315	3NPSA_R222A	-0,084	2,026	2,11
3HFMH_Y33W	1,720	1,816	0,096	3NPSA_R60CA	-1,062	2,667	3,729
3HFMH_Y50A	7,315	2,214	5,101	3NPSA_R60FA	0,138	1,259	1,121
3HFMH_Y50F	1,582	2,622	1,04	3NPSA_R87A	-0,151	1,292	1,443
3HFMH_Y50L	2,800	0,467	2,333	3NPSA_T150A	0,175	2,685	2,51
3HFMH_Y53A	3,195	0,471	2,724	3NPSA_T98A	0,723	0,548	0,175
3HFMH_Y53F	0,610	0,954	0,344	3NPSA_Y146A	1,774	0,683	1,091
3HFMH_Y53L	0,793	0,358	0,435	3NPSA_Y60GA	0,452	1,774	1,322
3HFMH_Y53W	0,694	1,926	1,232	3SGBI_A15C	0,041	0,818	0,777
3HFMH_Y58A	1,647	2,556	0,909	3SGBI_A15D	0,476	1,769	1,293
3HFMH_Y58F	0,357	1,004	0,647	3SGBI_A15E	0,671	0,392	0,279
3HFMH_Y58L	1,488	0,353	1,135	3SGBI_A15F	-1,179	2,652	3,831
3HFML_Y50A	4,555	1,751	2,804	3SGBI_A15G	2,484	0,084	2,4
3HFML_Y50F	2,353	1,698	0,655	3SGBI_A15H	-0,453	1,463	1,916
3HFML_Y50L	4,390	0,234	4,156	3SGBI_A15I	-0,453	1,063	1,516
3HFML_Y96A	2,705	2,078	0,627	3SGBI_A15K	2,489	0,927	1,562
3HFML_Y96F	1,401	0,686	0,715	3SGBI_A15L	0,169	1,445	1,276
3HFMY_Y20A	4,571	0,196	4,375	3SGBI_A15M	0,089	1,443	1,354
3HFMY_Y20F	-0,484	1,922	2,406	3SGBI_A15N	-0,089	2,882	2,971
3HFMY_Y20L	2,183	0,021	2,162	3SGBI_A15P	0,552	1,082	0,53
3NPSA_D217A	1,466	0,339	1,127	3SGBI_A15Q	0,430	2,764	2,334
3NPSA_D60AA	0,340	1,631	1,291	3SGBI_A15R	2,033	1,574	0,459
3NPSA_D60BA	1,067	1,139	0,072	3SGBI_A15S	0,910	2,578	1,668
3NPSA_D96A	1,506	0,801	0,705	3SGBI_A15T	0,331	2,167	1,836
3NPSA_E169A	0,615	0,590	0,025	3SGBI_A15V	-0,721	0,103	0,824
3NPSA_F60EA	0,219	0,544	0,325	3SGBI_A15W	-1,552	0,013	1,565
3NPSA_F94A	1,594	0,953	0,641	3SGBI_A15Y	-1,645	2,756	4,401
3NPSA_F97A	0,463	0,131	0,332	3SGBI_E19A	1,018	1,743	0,725

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3SGBI_E19C	1,173	0,446	0,727
3SGBI_E19D	0,552	3,000	2,448
3SGBI_E19F	1,941	1,661	0,28
3SGBI_E19G	2,103	2,306	0,203
3SGBI_E19H	0,525	0,548	0,023
3SGBI_E19I	-0,623	1,821	2,444
3SGBI_E19K	0,169	1,460	1,291
3SGBI_E19L	0,778	1,837	1,059
3SGBI_E19M	-0,189	1,806	1,995
3SGBI_E19N	1,101	0,520	0,581
3SGBI_E19P	2,273	0,179	2,094
3SGBI_E19Q	0,183	2,338	2,155
3SGBI_E19R	1,274	2,152	0,878
3SGBI_E19S	1,941	0,744	1,197
3SGBI_E19T	2,141	0,268	1,873
3SGBI_E19V	0,127	2,104	1,977
3SGBI_E19W	1,546	0,718	0,828
3SGBI_E19Y	0,778	0,276	0,502
3SGBI_G32A	1,292	2,228	0,936
3SGBI_G32C	2,609	1,311	1,298
3SGBI_G32D	1,623	2,873	1,25
3SGBI_G32E	1,971	1,432	0,539
3SGBI_G32F	3,019	1,542	1,477
3SGBI_G32H	2,735	0,412	2,323
3SGBI_G32I	4,088	0,436	3,652
3SGBI_G32K	2,522	0,944	1,578
3SGBI_G32L	2,655	0,409	2,246
3SGBI_G32M	2,803	2,215	0,588
3SGBI_G32N	2,273	0,849	1,424
3SGBI_G32P	1,096	0,597	0,499
3SGBI_G32Q	2,953	0,441	2,512
3SGBI_G32R	3,504	0,412	3,092
3SGBI_G32S	1,606	0,738	0,868
3SGBI_G32T	3,037	2,454	0,583
3SGBI_G32V	2,824	1,055	1,769
3SGBI_G32W	3,908	0,912	2,996
3SGBI_G32Y	3,093	1,394	1,699
3SGBI_K13A	-2,569	1,238	3,807
3SGBI_K13C	-0,584	2,004	2,588

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3SGBI_K13D	-0,623	0,308	0,931
3SGBI_K13E	0,032	1,529	1,497
3SGBI_K13F	-0,725	0,720	1,445
3SGBI_K13G	-0,725	2,014	2,739
3SGBI_K13H	-0,401	1,909	2,31
3SGBI_K13I	-1,343	0,502	1,845
3SGBI_K13L	-1,815	2,784	4,599
3SGBI_K13M	-1,165	0,415	1,58
3SGBI_K13N	-0,659	2,723	3,382
3SGBI_K13P	0,000	0,370	0,37
3SGBI_K13Q	-0,623	2,680	3,303
3SGBI_K13R	-0,692	0,290	0,982
3SGBI_K13S	-2,585	2,412	4,997
3SGBI_K13T	-1,947	2,191	4,138
3SGBI_K13V	-0,954	0,716	1,67
3SGBI_K13W	-0,013	0,009	0,022
3SGBI_K13Y	-0,584	0,898	1,482
3SGBI_L18A	2,970	1,460	1,51
3SGBI_L18C	-0,013	1,606	1,619
3SGBI_L18D	5,662	0,653	5,009
3SGBI_L18E	6,168	1,444	4,724
3SGBI_L18F	1,372	2,263	0,891
3SGBI_L18G	4,998	0,982	4,016
3SGBI_L18H	1,712	1,120	0,592
3SGBI_L18I	4,476	2,654	1,822
3SGBI_L18K	3,179	0,912	2,267
3SGBI_L18M	0,453	2,599	2,146
3SGBI_L18N	3,396	1,257	2,139
3SGBI_L18P	8,436	1,487	6,949
3SGBI_L18Q	2,543	0,664	1,879
3SGBI_L18R	3,363	1,858	1,505
3SGBI_L18S	4,154	1,301	2,853
3SGBI_L18T	3,201	0,931	2,27
3SGBI_L18V	3,037	0,751	2,286
3SGBI_L18W	1,863	0,963	0,9
3SGBI_L18Y	1,674	2,563	0,889
3SGBI_N28S	0,010	0,926	0,916
3SGBI_N36A	0,331	2,939	2,608
3SGBI_N36C	0,349	2,918	2,569

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3SGBI_N36D	0,910	2,771	1,861
3SGBI_N36E	0,962	2,104	1,142
3SGBI_N36F	-0,141	2,207	2,348
3SGBI_N36G	0,311	1,722	1,411
3SGBI_N36H	-0,125	1,990	2,115
3SGBI_N36I	0,198	1,784	1,586
3SGBI_N36K	0,579	2,948	2,369
3SGBI_N36L	0,740	0,389	0,351
3SGBI_N36M	0,141	2,099	1,958
3SGBI_N36P	0,579	0,331	0,248
3SGBI_N36Q	0,243	1,357	1,114
3SGBI_N36R	0,579	2,791	2,212
3SGBI_N36S	0,409	0,887	0,478
3SGBI_N36T	0,010	2,058	2,048
3SGBI_N36V	-0,242	1,632	1,874
3SGBI_N36W	0,525	1,986	1,461
3SGBI_N36Y	-0,109	2,032	2,141
3SGBI_P14A	-0,189	1,204	1,393
3SGBI_P14C	-0,327	2,922	3,249
3SGBI_P14D	-1,032	1,999	3,031
3SGBI_P14E	-0,623	1,229	1,852
3SGBI_P14F	-0,345	0,147	0,492
3SGBI_P14G	0,054	0,047	0,007
3SGBI_P14H	0,277	2,794	2,517
3SGBI_P14I	-0,024	1,831	1,855
3SGBI_P14K	0,114	0,771	0,657
3SGBI_P14L	-0,141	1,364	1,505
3SGBI_P14M	-0,500	1,507	2,007
3SGBI_P14N	-0,125	0,440	0,565
3SGBI_P14Q	-0,316	0,682	0,998
3SGBI_P14R	0,311	2,090	1,779
3SGBI_P14S	0,066	2,403	2,337
3SGBI_P14T	-0,109	2,745	2,854
3SGBI_P14V	-0,310	0,515	0,825
3SGBI_P14W	-0,500	0,341	0,841
3SGBI_P14Y	-0,401	0,160	0,561
3SGBI_R21A	0,054	2,776	2,722
3SGBI_R21C	0,552	1,400	0,848
3SGBI_R21D	0,638	1,674	1,036

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3SGBI_R21E	0,671	1,167	0,496
3SGBI_R21F	0,228	0,814	0,586
3SGBI_R21G	1,106	0,569	0,537
3SGBI_R21H	0,453	0,403	0,05
3SGBI_R21I	0,243	0,307	0,064
3SGBI_R21K	-0,277	0,328	0,605
3SGBI_R21L	0,228	0,716	0,488
3SGBI_R21M	0,127	2,598	2,471
3SGBI_R21N	0,331	1,842	1,511
3SGBI_R21P	7,676	1,416	6,26
3SGBI_R21Q	0,041	1,556	1,515
3SGBI_R21S	0,277	0,529	0,252
3SGBI_R21T	0,453	0,978	0,525
3SGBI_R21V	-0,024	0,305	0,329
3SGBI_R21W	0,311	1,590	1,279
3SGBI_R21Y	0,294	1,033	0,739
3SGBI_T17A	3,407	0,396	3,011
3SGBI_T17C	2,923	1,880	1,043
3SGBI_T17D	4,951	2,211	2,74
3SGBI_T17E	4,623	2,415	2,208
3SGBI_T17F	3,635	2,443	1,192
3SGBI_T17G	5,578	1,969	3,609
3SGBI_T17H	3,504	2,517	0,987
3SGBI_T17I	1,751	0,679	1,072
3SGBI_T17K	2,033	1,841	0,192
3SGBI_T17L	2,469	1,171	1,298
3SGBI_T17M	2,033	0,332	1,701
3SGBI_T17N	3,251	1,144	2,107
3SGBI_T17P	3,179	1,727	1,452
3SGBI_T17Q	2,937	2,202	0,735
3SGBI_T17R	1,971	2,830	0,859
3SGBI_T17S	2,500	1,541	0,959
3SGBI_T17V	2,258	2,218	0,04
3SGBI_T17W	3,037	0,464	2,573
3SGBI_T17Y	3,363	1,669	1,694
3SGBI_Y20A	1,941	0,954	0,987
3SGBI_Y20C	1,693	2,760	1,067
3SGBI_Y20D	2,861	1,938	0,923
3SGBI_Y20E	2,033	1,411	0,622

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
3SGBI_Y20F	0,169	1,241	1,072
3SGBI_Y20G	2,839	1,972	0,867
3SGBI_Y20H	1,558	0,135	1,423
3SGBI_Y20I	3,588	0,759	2,829
3SGBI_Y20K	3,429	0,393	3,036
3SGBI_Y20L	1,574	2,686	1,112
3SGBI_Y20M	2,001	0,250	1,751
3SGBI_Y20N	3,155	2,193	0,962
3SGBI_Y20P	6,269	1,213	5,056
3SGBI_Y20Q	3,277	2,489	0,788
3SGBI_Y20R	2,757	2,803	0,046
3SGBI_Y20S	2,033	0,551	1,482
3SGBI_Y20T	4,866	0,344	4,522
3SGBI_Y20V	4,417	0,067	4,35
3SGBI_Y20W	0,388	2,303	1,915
3TGKE_N194D	-0,170	2,316	2,486
1ACBI_L45D	6,859	1,670	5,189
1ACBI_L45E	6,639	1,400	5,239
1ACBI_L45G	6,111	1,991	4,12
1ACBI_L45I	4,285	2,755	1,53
1ACBI_L45P	6,941	1,804	5,137
1ACBI_L45S	5,015	0,753	4,262
1CBWI_F33A	0,143	1,354	1,211
1CBWI_G12A	0,685	1,510	0,825
1CBWI_G36A	0,963	0,487	0,476
1CBWI_G37A	0,820	0,906	0,086
1CBWI_I18A	1,414	0,069	1,345
1CBWI_I19A	0,143	0,625	0,482
1CBWI_K15A	2,110	2,921	0,811
1CBWI_K15D	5,265	0,730	4,535
1CBWI_K15E	3,902	0,885	3,017
1CBWI_K15F	-1,982	2,235	4,217
1CBWI_K15G	4,152	2,522	1,63
1CBWI_K15H	0,100	2,903	2,803
1CBWI_K15I	2,958	0,950	2,008
1CBWI_K15L	-1,596	2,043	3,639
1CBWI_K15M	-1,428	1,362	2,79
1CBWI_K15N	1,336	0,684	0,652
1CBWI_K15Q	0,308	2,332	2,024

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1CBWI_K15R	-0,620	0,607	1,227
1CBWI_K15S	3,422	2,039	1,383
1CBWI_K15T	2,106	2,374	0,268
1CBWI_K15V	2,155	2,754	0,599
1CBWI_K15W	-2,458	0,646	3,104
1CBWI_K15Y	-2,639	0,971	3,61
1CBWI_K46A	0,143	2,914	2,771
1CBWI_P13A	-0,056	2,658	2,714
1CBWI_R17A	0,553	1,549	0,996
1CBWI_R20A	0,354	2,768	2,414
1CBWI_R39A	0,222	0,196	0,026
1CBWI_T11A	0,222	1,674	1,452
1CBWI_V34A	0,051	1,784	1,733
1CBWI_Y35A	0,884	0,898	0,014
1DVFB_D100A	2,788	2,139	0,649
1DVFB_D54A	4,278	2,002	2,276
1DVFB_D58A	1,599	0,592	1,007
1DVFD_D52A	1,681	2,952	1,271
1DVFB_E98A	4,184	0,984	3,2
1DVFA_H30A	1,648	2,064	0,416
1DVFD_H33A	1,859	1,368	0,491
1DVFD_I97A	2,680	0,649	2,031
1DVFD_K30A	1,003	2,140	1,137
1DVFB_N56A	1,162	2,752	1,59
1DVFD_N54A	1,859	1,340	0,519
1DVFD_Q100A	1,628	1,443	0,185
1DVFB_R99A	1,873	0,195	1,678
1DVFD_R100BA	4,088	2,646	1,442
1DVFA_S93A	1,162	0,646	0,516
1DVFB_T30A	0,907	1,290	0,383
1DVFA_W92A	0,340	1,514	1,174
1DVFB_W52A	4,130	0,699	3,431
1DVFA_Y32A	2,029	1,955	0,074
1DVFA_Y49A	1,678	1,678	0
1DVFA_Y50A	0,687	2,526	1,839
1DVFB_Y101F	2,011	1,595	0,416
1DVFB_Y32A	1,830	0,398	1,432
1DVFC_Y49A	1,859	0,205	1,654
1DVFD_Y98A	4,736	2,831	1,905

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1EAWA_D217A	2,227	1,541	0,686	1FFWB_F214A	3,642	2,694	0,948
1EAWA_D60AA	-0,171	2,210	2,381	1FFWB_H181A	0,034	2,478	2,444
1EAWA_D60BA	1,501	0,430	1,071	1FFWB_I216A	0,427	1,309	0,882
1EAWA_D96A	0,654	1,517	0,863	1FFWA_T112I	0,562	2,108	1,546
1EAWA_E169A	0,703	0,298	0,405	1FFWA_T87I	-0,319	0,613	0,932
1EAWA_F60EA	-0,428	2,936	3,364	1FFWA_V108M	1,130	2,742	1,612
1EAWA_F94A	0,728	1,785	1,057	1FFWA_Y106W	0,713	2,870	2,157
1EAWA_F97A	0,891	0,120	0,771	1GC1C_D10A	0,000	0,123	0,123
1EAWA_H143A	-0,014	1,243	1,257	1GC1C_D53A	0,302	1,008	0,706
1EAWA_I41A	-0,822	2,854	3,676	1GC1C_D56A	-0,070	1,669	1,739
1EAWA_I60A	-0,194	1,543	1,737	1GC1C_D63A	-0,319	2,322	2,641
1EAWA_K224A	0,476	2,154	1,678	1GC1C_D88A	-0,070	0,603	0,673
1EAWA_L153A	0,502	0,138	0,364	1GC1C_E77A	0,562	2,049	1,487
1EAWA_N95A	0,308	0,229	0,079	1GC1C_E85A	1,322	2,437	1,115
1EAWA_Q145A	0,306	0,733	0,427	1GC1C_E87A	0,218	0,037	0,181
1EAWA_Q174A	0,564	2,976	2,412	1GC1C_E91A	-0,128	1,874	2,002
1EAWA_Q175A	-0,133	0,681	0,814	1GC1C_E92A	0,016	0,259	0,243
1EAWA_Q221AA	0,144	0,479	0,335	1GC1C_H27A	0,282	1,672	1,39
1EAWA_Q38A	-0,518	0,351	0,869	1GC1C_K1A	0,062	1,583	1,521
1EAWA_R222A	-0,088	0,317	0,405	1GC1C_K21A	-0,128	2,898	3,026
1EAWA_R60CA	0,587	0,532	0,055	1GC1C_K22A	0,240	0,742	0,502
1EAWA_R60FA	0,231	1,708	1,477	1GC1C_K29A	0,536	1,015	0,479
1EAWA_R87A	-0,150	0,945	1,095	1GC1C_K2A	-0,017	2,848	2,865
1EAWA_T150A	0,089	2,601	2,512	1GC1C_K35A	0,322	2,197	1,875
1EAWA_T98A	0,254	1,171	0,917	1GC1C_K46A	1,429	2,442	1,013
1EAWA_Y146A	0,502	1,049	0,547	1GC1C_K50A	0,047	2,618	2,571
1EAWA_Y60GA	-0,079	1,735	1,814	1GC1C_K72A	-0,017	2,763	2,78
1EFNA_I96A	1,449	2,527	1,078	1GC1C_K75A	0,158	2,964	2,806
1EFNA_T97H	1,242	1,429	0,187	1GC1C_K7A	0,997	0,528	0,469
1FFWA_A90V	0,091	2,563	2,472	1GC1C_K8A	0,105	2,750	2,645
1FFWB_C213A	0,204	2,530	2,326	1GC1C_K90A	0,047	0,165	0,118
1FFWA_D13K	0,047	1,961	1,914	1GC1C_L44A	1,055	1,574	0,519
1FFWB_D202A	-0,074	2,216	2,29	1GC1C_L51A	1,233	1,376	0,143
1FFWB_D207A	0,096	2,196	2,1	1GC1C_N30A	0,170	2,158	1,988
1FFWA_E117K	0,713	1,331	0,618	1GC1C_N32A	0,182	0,661	0,479
1FFWA_E93K	0,820	0,069	0,751	1GC1C_N39A	0,465	2,832	2,367
1FFWB_E171A	0,716	2,350	1,634	1GC1C_N52A	0,708	0,175	0,533
1FFWB_E178A	0,638	2,543	1,905	1GC1C_N66A	-0,034	1,379	1,413
1FFWA_F111V	6,705	2,665	4,04	1GC1C_N73A	-0,108	0,369	0,477

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1GC1C_Q20A	-0,017	0,709	0,726
1GC1C_Q25A	0,032	2,438	2,406
1GC1C_Q33A	0,105	1,064	0,959
1GC1C_Q40A	-0,410	1,005	1,415
1GC1C_Q64A	0,442	1,834	1,392
1GC1C_Q89A	0,170	0,155	0,015
1GC1C_Q89L	-0,311	1,272	1,583
1GC1C_Q94A	-0,108	2,894	3,002
1GC1C_R58A	0,132	1,654	1,522
1GC1C_R59A	1,175	2,589	1,414
1GC1C_S19A	0,000	1,981	1,981
1GC1C_S23A	0,292	2,138	1,846
1GC1C_S31A	0,105	1,749	1,644
1GC1C_S42A	0,000	2,328	2,328
1GC1C_S49A	0,610	0,396	0,214
1GC1C_S60A	-0,088	0,995	1,083
1GC1C_T11A	0,000	0,789	0,789
1GC1C_T15A	0,322	0,957	0,635
1GC1C_T17A	-0,128	2,419	2,547
1GC1C_T45A	-0,149	0,226	0,375
1GC1C_V86A	-0,070	2,872	2,942
1GC1C_Y82A	1,289	1,681	0,392
1GRNA_D38E	0,389	2,789	2,4
1H9DB_G61A	2,075	2,903	0,828
1H9DB_L103A	0,939	0,149	0,79
1H9DB_N104A	2,302	1,037	1,265
1H9DB_Q67A	1,363	1,736	0,373
1H9DB_R3A	1,161	2,570	1,409
1H9DB_V4A	1,401	1,642	0,241
1HE8A_K223V	0,475	0,501	0,026
1JRHH_D54A	1,885	0,578	1,307
1JRHH_D55A	1,665	1,389	0,276
1JRHH_D56A	1,853	2,208	0,355
1JRHL_D28A	0,434	0,874	0,44
1JRHI_E45Q	0,100	0,748	0,648
1JRHI_E55A	-0,566	1,116	1,682
1JRHI_E55P	-3,786	0,338	4,124
1JRHL_E27A	0,542	1,339	0,797
1JRHH_F98A	0,000	2,684	2,684

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1JRHI_G50A	4,447	0,402	4,045
1JRHH_H100BA	1,697	2,958	1,261
1JRHI_K47A	3,712	1,585	2,127
1JRHI_K47M	3,290	2,050	1,24
1JRHI_K52A	3,384	2,493	0,891
1JRHI_K52M	4,908	0,317	4,591
1JRHI_K98A	0,312	0,105	0,207
1JRHI_N48A	0,170	2,575	2,405
1JRHI_N48Q	0,534	0,561	0,027
1JRHI_N53A	4,297	1,081	3,216
1JRHI_N79A	-0,199	0,720	0,919
1JRHH_R95A	0,542	0,356	0,186
1JRHI_R84A	0,393	2,623	2,23
1JRHI_S54A	0,377	0,046	0,331
1JRHL_S93A	-0,650	1,359	2,009
1JRHI_T14V	-0,026	0,659	0,685
1JRHL_T94A	0,385	2,218	1,833
1JRHI_V51A	1,678	0,112	1,566
1JRHH_W52A	2,684	2,908	0,224
1JRHH_W53A	2,420	2,205	0,215
1JRHI_W56F	-0,252	2,300	2,552
1JRHI_W56Y	0,298	1,883	1,585
1JRHI_W82A	4,426	2,516	1,91
1JRHI_W82F	1,102	0,929	0,173
1JRHI_W82Y	1,187	0,098	1,089
1JRHL_W92A	2,817	0,705	2,112
1JRHL_W96A	1,665	0,231	1,434
1JRHH_Y99A	1,060	2,039	0,979
1JRHH_Y32A	1,432	0,500	0,932
1JRHH_Y58A	1,255	2,386	1,131
1JRHI_Y49A	3,525	1,428	2,097
1JRHI_Y49F	0,908	2,983	2,075
1JRHL_Y30A	1,108	1,633	0,525
1JRHL_Y91A	0,580	2,214	1,634
1LFDA_D51A	-0,578	0,169	0,747
1LFDA_D51K	-1,085	0,206	1,291
1LFDA_D56A	-0,279	0,521	0,8
1LFDA_D58K	-0,906	1,987	2,893
1LFDA_D94K	-1,146	2,371	3,517

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro	Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
1LFDA_E57A	-0,246	2,894	3,14	1VFBC_Q121A	2,875	1,256	1,619
1LFDA_E57K	-0,284	1,359	1,643	1VFBB_R99A	-0,100	1,088	1,188
1LFDA_K32A	1,325	1,111	0,214	1VFBC_R125A	1,836	1,670	0,166
1LFDA_K48A	0,258	1,479	1,221	1VFBA_S93A	0,343	2,085	1,742
1LFDA_K52A	1,178	1,064	0,114	1VFBC_S24A	0,850	2,417	1,567
1LFDA_L23K	-0,015	0,880	0,895	1VFBB_T30A	-0,056	2,829	2,885
1LFDA_L55K	-0,570	1,209	1,779	1VFBC_T118A	0,765	0,895	0,13
1LFDA_M30K	-0,920	0,405	1,325	1VFBC_V120A	0,916	1,988	1,072
1LFDA_N27K	0,403	1,399	0,996	1VFBA_W92A	3,036	1,257	1,779
1LFDA_N54K	-1,167	2,779	3,946	1VFBB_W52A	0,640	0,317	0,323
1LFDA_N92K	-0,562	0,698	1,26	1VFBA_Y32A	1,527	0,818	0,709
1LFDA_R20A	1,135	1,154	0,019	1VFBA_Y49A	0,797	0,135	0,662
1LFDA_S22K	-0,118	0,677	0,795	1VFBA_Y50A	0,456	1,224	0,768
1LFDA_Y93K	0,077	2,924	2,847	1VFBB_Y101F	1,269	2,714	1,445
1NMBH_D56E	2,952	0,587	2,365	1VFBB_Y32A	0,791	0,899	0,108
1NMBH_D56N	2,952	0,163	2,789	1VFBC_Y23A	0,410	2,352	1,942
1NMBL_L94V	0,899	2,055	1,156	2B42A_H374A	1,637	0,064	1,573
1NMBL_T93F	-0,024	0,786	0,81	2B42A_H374K	2,565	2,262	0,303
1NMBL_T93W	0,184	0,763	0,579	2B42A_H374Q	1,075	2,117	1,042
1NMBH_Y99A	2,139	1,459	0,68	2FTLI_G12A	4,388	0,800	3,588
1NMBH_Y100AF	1,418	2,004	0,586	2FTLI_G36A	2,212	0,649	1,563
1NMBL_Y32F	0,466	2,644	2,178	2FTLI_I18A	5,016	0,846	4,17
1REWC_Q86A	2,656	1,209	1,447	2FTLI_K15A	10,387	2,284	8,103
1SMFI_E16A	1,012	1,112	0,1	2FTLI_K15D	11,490	2,879	8,611
1SMFI_I13A	3,508	1,932	1,576	2FTLI_K15E	9,414	1,545	7,869
1SMFI_S12A	1,898	0,326	1,572	2FTLI_K15F	7,019	2,087	4,932
1SMFI_T10A	2,044	1,832	0,212	2FTLI_K15G	12,338	2,799	9,539
1VFBB_D100A	3,003	2,184	0,819	2FTLI_K15H	8,772	1,736	7,036
1VFBB_D54A	0,814	2,424	1,61	2FTLI_K15I	11,159	0,284	10,875
1VFBB_D58A	-0,207	2,566	2,773	2FTLI_K15L	8,854	2,210	6,644
1VFBC_D119A	0,952	2,560	1,608	2FTLI_K15M	7,684	1,201	6,483
1VFBC_D18A	0,340	2,103	1,763	2FTLI_K15N	8,024	2,372	5,652
1VFBB_E98A	1,156	2,832	1,676	2FTLI_K15Q	9,360	2,258	7,102
1VFBA_H30A	0,844	1,851	1,007	2FTLI_K15S	7,715	1,174	6,541
1VFBC_I124A	1,231	0,761	0,47	2FTLI_K15T	10,586	2,095	8,491
1VFBC_K116A	0,713	1,252	0,539	2FTLI_K15V	11,743	0,955	10,788
1VFBC_L129A	0,171	0,978	0,807	2FTLI_K15W	8,654	1,488	7,166
1VFBB_N56A	0,178	1,133	0,955	2FTLI_K15Y	6,849	1,576	5,273
1VFBC_N19A	0,396	2,731	2,335	2G2UA_D104K	-0,452	0,417	0,869

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
2G2UB_E31A	0,650	0,358	0,292
2G2UB_E73A	-1,977	0,537	2,514
2G2UB_E73M	-3,532	1,609	5,141
2G2UB_F142A	0,276	0,437	0,161
2G2UB_F36A	2,761	0,359	2,402
2G2UB_G141A	-0,413	0,482	0,895
2G2UB_G48A	-0,426	1,332	1,758
2G2UB_H148A	1,118	2,214	1,096
2G2UB_H41A	1,715	1,074	0,641
2G2UB_K74A	-0,217	1,726	1,943
2G2UB_R144A	-0,342	2,411	2,753
2G2UB_R160A	0,669	0,877	0,208
2G2UB_S113A	-0,612	0,153	0,765
2G2UB_S130K	-1,468	1,133	2,601
2G2UB_S35A	-0,950	2,207	3,157
2G2UB_S39A	-0,955	1,635	2,59
2G2UB_S71A	-0,512	1,555	2,067
2G2UB_W112A	0,958	2,871	1,913
2G2UB_W150A	1,783	0,687	1,096
2G2UB_W162A	0,531	0,824	0,293
2G2UB_Y143A	-1,845	2,059	3,904
2G2UB_Y50A	-2,073	1,821	3,894
2G2UB_Y51A	-0,628	2,929	3,557
2G2UB_Y53A	2,299	0,182	2,117
2J12A_S299A	0,952	1,122	0,17
2J1KC_G370D	0,184	2,066	1,882
2J1KC_R515A	0,051	0,962	0,911
2J1KC_R384A	0,764	1,611	0,847
2O3BB_D75E	5,432	2,172	3,26
2O3BB_D75N	5,897	0,025	5,872
2O3BB_E24A	5,469	2,610	2,859
2O3BB_E24D	0,599	0,974	0,375
2O3BB_E24Q	5,392	0,040	5,352
2O3BB_Q74A	3,229	0,279	2,95
2O3BB_W76A	4,069	1,274	2,795
2OOBA_G941S	0,235	2,533	2,298
2OOBA_K935E	-0,178	2,841	3,019
2PCBA_A193F	0,733	0,608	0,125
2PCBA_D34N	0,820	0,517	0,303

Código da mutação	ΔΔG Experimental	ΔΔG Predito	Erro
2PCBA_E290N	0,870	0,761	0,109
2PCBA_E291Q	-0,056	2,255	2,311
2PCBA_E32Q	0,605	2,992	2,387
2PCBA_E35Q	0,675	2,196	1,521
2PCCB_A81G	1,900	2,004	0,104
2PCCA_D34A	-0,896	0,430	1,326
2PCCA_E290A	6,196	2,138	4,058
2PCCB_K72A	0,303	0,399	0,096
2PCCB_K87A	0,901	1,582	0,681
2PCCA_V197A	2,100	0,713	1,387
4CPAI_P36G	2,593	1,820	0,773
4CPAI_V38A	2,323	0,959	1,364
4CPAI_V38F	0,000	2,337	2,337
4CPAI_V38G	3,731	2,807	0,924
4CPAI_V38I	0,199	1,025	0,826
4CPAI_V38L	1,278	2,702	1,424
4CPAI_Y37F	0,000	1,434	1,434
4CPAI_Y37G	0,501	2,065	1,564

Anexo B

Atributos utilizados no processo de regressão

A Tabela a seguir apresenta quais atributos são destinados para cada regressor.

Atributo	**Origem**	**SVR**	**Processo Gaussiano**	**KNN**	**M5P**
Entropia do complexo	Mutante	X		X	X
	Selvagem	X	X	X	X
Quantidade de pontes de hidrogênio	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Diâmetro do grafo	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Número de resíduso polares	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Closeness	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Ecentricidades	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Impureza de vértices	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Entropia do complexo	Mutante		X	X	X
	Selvagem	X	X	X	X
Vértices terminais	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Van der Waals	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Soma dos altovalores	Mutante	X	X	X	X

Atributo	**Origem**	**SVR**	**Processo Gaussiano**	**KNN**	**M5P**
	Selvagem	X	X	X	X
Quantidade de pontes de hidrogênio na cadeia principal	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Impureza de arestas	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Resíduos apolares	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Quantidade de resíduos na interface	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Quantidade de contatos aromáticos	Mutante	X		X	X
	Selvagem	X	X	X	X
Betweenness	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Quantidade de contatos atrativos	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Interações eletrostáticas contabilizadas pelo FoldX	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Número de vértices do grafo	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Energia do complexo	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Grau	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Maior altovalor	Mutante	X	X	X	X
	Selvagem		X	X	X
Segundo maior altovalor	Mutante		X	X	X
	Selvagem	X	X	X	X
Quantidade de pontes de hidrogênio na cadeia lateral	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Área de contato da interface	Mutante	X	X	X	X
	Selvagem		X	X	X
Arestas	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Quantidade de pontes dissulfeto computadas pelo FoldX	Mutante		X	X	X
	Selvagem	X	X	X	X

Atributo	**Origem**	**SVR**	**Processo Gaussiano**	**KNN**	**M5P**
Quantidade de pontos centrais	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Energia da cadeia lateral	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Energia da cadeia principal	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Authority score	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Entropia na conformação da cadeia principal	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Entropia na conformação da cadeia lateral	Mutante	X	X	X	X
	Selvagem	X	X	X	X
eletrostatic	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Quantidade de pontes de hidrogênio do grafo	Mutante		X	X	X
	Selvagem	X	X	X	X
Quantidade de contatos não hidrofóbicos do grafo	Mutante	X	X	X	X
	Selvagem		X	X	X
Quantidade de contatos não classificados do grafo	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Conflitos de Van der Waals	Mutante	X	X	X	X
	Selvagem	X	X	X	X
Quantidade de contatos repulsivos do grafo	Mutante	X	X	X	X
	Selvagem	X	X	X	X

Referências Bibliográficas

[Abagyan & Totrov, 1994] Abagyan, R. & Totrov, M. (1994). Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *Journal of molecular biology*, 235(3):983--1002.

[Adamic, 1999] Adamic, L. A. (1999). The small world web. Em *Research and Advanced Technology for Digital Libraries*, pp. 443--452. Springer.

[Aittokallio & Schwikowski, 2006] Aittokallio, T. & Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, 7(3):243--255.

[Alpaydin, 2007] Alpaydin, E. (2007). Combining pattern classifiers: Methods and algorithms (kuncheva, li; 2004)[book review]. *IEEE Transactions on Neural Networks*, 18(3):964--964.

[Bank, 1971] Bank, P. D. (1971). Protein data bank. *Nature New Biol*, 233:223.

[Barabási & Albert, 1999] Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509--512.

[Barabasi & Oltvai, 2004] Barabasi, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101--113.

[Barrat & Weigt, 2000] Barrat, A. & Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547--560.

[Bassett & Bullmore, 2016] Bassett, D. S. & Bullmore, E. T. (2016). Small-world brain networks revisited. *The Neuroscientist*, p. 1073858416667720.

[Bates & Watts, 2008] Bates, D. M. & Watts, D. G. (2008). *Nonlinear Regression: Iterative Estimation and Linear Approximations*. John Wiley & Sons, Inc.

[Becker et al., 1988] Becker, R. A.; Chambers, J. M. & Wilks, A. R. (1988). The new s language. *Pacific Grove, Ca.: Wadsworth & Brooks, 1988*, 1.

[Berg et al., 2014] Berg, J. M.; Tymoczko, J. L.; Stryer, L. & Gatto Jr, G. J. (2014). Bioquímica. Em *Bioquímica*. Guanabara koogan.

[Bergholdt et al., 2012] Bergholdt, R.; Brorsson, C.; Palleja, A.; Berchtold, L. A.; Fløyel, T.; Bang-Berthelsen, C. H.; Frederiksen, K. S.; Jensen, L. J.; Størling, J. & Pociot, F. (2012). Identification of novel type 1 diabetes candidate genes by integrating genome-wide association data, protein-protein interactions, and human pancreatic islet gene expression. *Diabetes*, 61(4):954–962.

[Boccaletti et al., 2006] Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M. & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4):175--308.

[Boser et al., 1992] Boser, B. E.; Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Em *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144--152. ACM.

[Brender & Zhang, 2015] Brender, J. R. & Zhang, Y. (2015). Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLOS Computational Biology*, 11(10):1–25.

[Broder et al., 2000] Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A. & Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33(1–6):309 – 320.

[Brooks et al., 1983] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. a. & Karplus, M. (1983). Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187--217.

[Bullock et al., 2000] Bullock, A. N.; Henckel, J.; Fersht, A. R. et al. (2000). Quantitative analysis of residual folding and dna binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene*, 19(10):1245--1256.

[Chan & Paelinckx, 2008] Chan, J. C.-W. & Paelinckx, D. (2008). Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6):2999--3011.

[Coifman & Wickerhauser, 1992] Coifman, R. R. & Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on information theory*, 38(2):713--718.

[Cortes & Vapnik, 1995] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273--297.

[Crone et al., 2006] Crone, S. F.; Guajardo, J. & Weber, R. (2006). A study on the ability of support vector regression and neural networks to forecast basic time series patterns. Em *Artificial Intelligence in Theory and Practice*, pp. 149--158. Springer.

[De Las Rivas & Fontanillo, 2010] De Las Rivas, J. & Fontanillo, C. (2010). Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807.

[Dehouck et al., 2013] Dehouck, Y.; Kwasigroch, J. M.; Rooman, M. & Gilis, D. (2013). Beatmusic: prediction of changes in protein–protein binding affinity on mutations. 41(W1):W333–W339.

[Deng & Runger, 2012] Deng, H. & Runger, G. (2012). Feature selection via regularized trees. Em *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1--8. IEEE.

[Deng et al., 2016] Deng, Z.; Zhu, X.; Cheng, D.; Zong, M. & Zhang, S. (2016). Efficient knn classification algorithm for big data. *Neurocomputing*, 195:143--148.

[Ekman et al., 2005] Ekman, D.; Björklund, Å. K.; Frey-Skött, J. & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of molecular biology*, 348(1):231--243.

[Eldén, 2006] Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15:327--384.

[Eldén, 2007] Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*, volume 4. SIAM.

[Engin et al., 2012] Engin, H. B.; Keskin, O.; Nussinov, R. & Gursoy, A. (2012). A strategy based on protein–protein interface motifs may help in identifying drug off-targets. *Journal of chemical information and modeling*, 52(8):2273--2286.

[Erdös & Rényi, 1970] Erdös, P. & Rényi, A. (1970). On a new law of large numbers. *Journal d'analyse mathématique*, 23(1):103--111.

[Fernandes et al., 2013] Fernandes, P.; O'Kelly, M.; Papadopoulos, C. & Sales, A. (2013). Analysis of exponential reliable production lines using kronecker descriptors. *International Journal of Production Research*, 51(14):4240--4257.

[Fornili et al., 2013] Fornili, A.; Pandini, A.; Lu, H.-C. & Fraternali, F. (2013). Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *Journal of chemical theory and computation*, 9(11):5127--5147.

[Freeman, 1979] Freeman, L. C. (1978–1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239.

[Gaines & Andreae, 1966] Gaines, B. R. & Andreae, J. H. (1966). A learning machine in the context of the general control problem. Em *Proceedings of the 3rd Congress of the International Federation for Automatic Control*. Butterworths London.

[Galiza Neto & Pitombeira, 2003] Galiza Neto, G. C. d. & Pitombeira, M. d. S. (2003). Aspectos moleculares da anemia falciforme. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, 39:51 – 56.

[Gantz & Reinsel, 2011] Gantz, J. & Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, 1142(2011):1--12.

[Girvan & Newman, 2002] Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[Golumbic, 2004] Golumbic, M. C. (2004). *Algorithmic Graph Theory and Perfect Graphs (Annals of Discrete Mathematics, Vol 57)*. North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands.

[Gonçalves et al., 2015] Gonçalves, W. R.; Gonçalves-Almeida, V. M.; Arruda, A. L.; Meira, W.; da Silveira, C. H.; Pires, D. E. & de Melo-Minardi, R. C. (2015). Pdbest: a user–friendly platform for manipulating and enhancing protein structures. *Bioinformatics*, p. btv223.

[Hauke & Kossowski, 2011] Hauke, J. & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87.

[Hiromoto, 2016] Hiromoto, R. E. (2016). Parallelism and complexity of a small-world network model. *International Journal of Computing*, 15(2):72--83.

[Hoerger, 2013] Hoerger, M. (2013). Zh: An updated version of steiger's z and web-based calculator for testing the statistical significance of the difference between dependent correlations. *Retrieved March*, 1:2014.

[Hubbard & Thornton, 1993] Hubbard, S. J. & Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1).

[J. & M., 1993] J., H. S. & M., T. J. (1993). Naccess v2.1.1.

[Jacobs et al., 2001] Jacobs, D. J.; Rader, A. J.; Kuhn, L. A. & Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins: Structure, Function, and Bioinformatics*, 44(2):150--165.

[Janin et al., 2003] Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I. & Wodak, S. J. (2003). Capri: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2--9.

[Jones & Thornton, 1996] Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13--20.

[Keskin et al., 2008] Keskin, O.; Gursoy, A.; Ma, B. & Nussinov, R. (2008). Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical reviews*, 108(4):1225--1244.

[Klug et al., 2009] Klug, W. S.; Cummings, M. R.; Spencer, C. A. & Palladino, M. A. (2009). *Conceitos de Genética*. Artmed Editora.

[Lee & Preacher, 2013] Lee, I. & Preacher, K. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [computer software]. *Retrieved January*, 7:2015.

[Li et al., 2012] Li, G.; Semerci, M.; Yener, B. & Zaki, M. J. (2012). Effective graph classification based on topological and label attributes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(4):265--283.

[Li et al., 2016] Li, M.; Simonetti, F. L.; Goncearenco, A. & Panchenko, A. R. (2016). Mutabind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic acids research*, 44(W1):W494--W501.

[Li et al., 2001] Li, W.-H.; Gu, Z.; Wang, H. & Nekrutenko, A. (2001). Evolutionary analyses of the human genome. *Nature*, 409(6822):847--849.

[Lin, 1991] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145--151.

[Lu et al., 2008] Lu, M.; Dousis, A. D. & Ma, J. (2008). Opus-psp: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of molecular biology*, 376(1):288--301.

[Mazar, 2008] Mazar, A. P. (2008). Urokinase plasminogen activator receptor choreographs multiple ligand interactions: implications for tumor progression and therapy. *Clinical Cancer Research*, 14(18):5649--5655.

[Metz et al., 2011] Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K.-H. & Gohlke, H. (2011). Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein–protein interface. *Journal of chemical information and modeling*, 52(1):120--133.

[Michalski et al., 2013] Michalski, R. S.; Carbonell, J. G. & Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

[Miller et al., 1987] Miller, S.; Janin, J.; Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641--656.

[Moal & Fernández-Recio, 2012] Moal, I. H. & Fernández-Recio, J. (2012). Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20):2600--2607.

[Moal & Fernández-Recio, 2012] Moal, I. H. & Fernández-Recio, J. (2012). Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. 28(20):2600–2607.

[Moretti et al., 2013] Moretti, R.; Fleishman, S. J.; Agius, R.; Torchala, M.; Bates, P. A.; Kastritis, P. L.; Rodrigues, J. P.; Trellet, M.; Bonvin, A. M.; Cui, M. et al. (2013). Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 81(11):1980--1987.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Nelson & Cox, 2008] Nelson, D. L. & Cox, M. M. (2008). *Lehninger Principles of Biochemistry*. W. H. Freeman, fifth edition edição.

[O'Roak et al., 2012] O'Roak, B. J.; Vives, L.; Girirajan, S.; Karakoc, E.; Krumm, N.; Coe, B. P.; Levy, R.; Ko, A.; Lee, C.; Smith, J. D. et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246--250.

[Phillips et al., 2005] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L. & Schulten, K. (2005). Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781--1802.

[Pires et al., 2013] Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Campos, F. F. & Meira, W. (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855--861.

[Pires et al., 2014] Pires, D. E. V.; Ascher, D. B. & Blundell, T. L. (2014). mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342.

[Potapov et al., 2009] Potapov, V.; Cohen, M. & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering Design and Selection*, 22(9):553--560.

[Puth et al., 2015] Puth, M.-T.; Neuhäuser, M. & Ruxton, G. D. (2015). Effective use of spearman's and kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102:77--84.

[Qi et al., 2006] Qi, Y.; Bar-Joseph, Z. & Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, 63(3):490--500.

[Quinlan et al., 1992] Quinlan, J. R. et al. (1992). Learning with continuous classes. Em *5th Australian joint conference on artificial intelligence*, volume 92, pp. 343--348. Singapore.

[Rahimikhoob et al., 2013] Rahimikhoob, A.; Asadi, M. & Mashal, M. (2013). A comparison between conventional and m5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. *Water resources management*, 27(14):4815--4826.

[Ranshous et al., 2015] Ranshous, S.; Shen, S.; Koutra, D.; Harenberg, S.; Faloutsos, C. & Samatova, N. F. (2015). Anomaly detection in dynamic networks: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3):223--247.

[Rao & Rao, 2014] Rao, M. B. & Rao, C. (2014). Bayesian networks. *Handbook of Statistics: Computational Statistics with R*, 32:357.

[Rujirapipat et al., 2017] Rujirapipat, S.; McGarry, K. & Nelson, D. (2017). Bioinformatic analysis using complex networks and clustering proteins linked with alzheimer's disease. Em *Advances in Computational Intelligence Systems*, pp. 219--230. Springer.

[Samatova et al., 2013] Samatova, N. F.; Hendrix, W.; Jenkins, J.; Padmanabhan, K. & Chakraborty, A. (2013). *Practical Graph Mining with R*. Chapman & Hall/CRC.

[Schölkopf et al., 1998] Schölkopf, B.; Smola, A. & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299--1319.

[Schroff et al., 2008] Schroff, F.; Criminisi, A. & Zisserman, A. (2008). Object class segmentation using random forests. Em *BMVC*, pp. 1--10.

[Schymkowitz et al., 2005a] Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F. & Serrano, L. (2005a). The foldx web server: an online force field. *Nucleic acids research*, 33(suppl 2):W382--W388.

[Schymkowitz et al., 2005b] Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F. & Serrano, L. (2005b). The foldx web server: an online force field. *Nucleic acids research*, 33(suppl 2):W382--W388.

[Selzer et al., 2000] Selzer, T.; Albeck, S. & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nature Structural & Molecular Biology*, 7(7):537--541.

[Sobolev et al., 1999] Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332.

[Steiger, 1980] Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

[Stuart & Peter, 2016] Stuart, R. & Peter, N. (2016). Artificial intelligence-a modern approach 3rd ed.

[Svetnik et al., 2003] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P. & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947--1958.

[Touw et al., 2015] Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A.; Krieger, E.; Joosten, R. P. & Vriend, G. (2015). A series of pdb-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364--D368.

[Vagenende et al., 2009] Vagenende, V.; Yap, M. G. & Trout, B. L. (2009). Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. *Biochemistry*, 48(46):11084--11096.

[Van Durme et al., 2011] Van Durme, J.; Delgado, J.; Stricher, F.; Serrano, L.; Schymkowitz, J. & Rousseau, F. (2011). A graphical interface for the foldx forcefield. *Bioinformatics*, 27(12):1711.

[Vapnik & Lerner, 1963] Vapnik, V. & Lerner, A. (1963). Generalized portrait method for pattern recognition. *Automation and Remote Control*, 24(6):774--780.

[Vendruscolo et al., 2002] Vendruscolo, M.; Dokholyan, N.; Paci, E. & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review E*, 65(6):061910.

[Vishwanathan et al., 2010] Vishwanathan, S. V. N.; Schraudolph, N. N.; Kondor, R. & Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201--1242.

[Vázquez et al., 2015] Vázquez, M.; Valencia, A. & Pons, T. (2015). Structure-ppi: a module for the annotation of cancer-related single-nucleotide variants at protein–protein interfaces. *Bioinformatics*, 31(14):2397–2399.

[Watts & Strogatz, 1998] Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440--442.

[Wu et al., 2009] Wu, C.-H.; Tzeng, G.-H. & Lin, R.-H. (2009). A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications*, 36(3):4725--4735.

[Wu et al., 2010] Wu, G.; Feng, X. & Stein, L. (2010). Research a human functional protein interaction network and its application to cancer data analysis. *Genome Biol*, 11:R53.

[Zaki & Wagner Meira, 2014] Zaki, M. J. & Wagner Meira, J. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

[Zhang & Wiemann, 2009] Zhang, J. D. & Wiemann, S. (2009). Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, 25(11):1470--1471.

[Zhang et al., 2013] Zhang, T.; Cao, J.; Chen, Y.; Cuthbert, L. & Elkashlan, M. (2013). A small world network model for energy efficient wireless networks. *IEEE Communications Letters*, 17(10):1928--1931.

[Zhang et al., 2017] Zhang, W.; Ben-David, M. & Sidhu, S. S. (2017). Engineering cell signaling modulators from native protein–protein interactions. *Current opinion in structural biology*, 45:25--35.

[Zhao et al., 2014] Zhao, N.; Han, J. G.; Shyu, C.-R. & Korkin, D. (2014). Determining effects of non-synonymous snps on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput Biol*, 10(5):e1003592.

[Zhou & Skolnick, 2011] Zhou, H. & Skolnick, J. (2011). Goap: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal*, 101(8):2043--2052.