Renan Martinez da Luz

Descoberta de Conhecimento com Auxílio da Inteligência Humana: um estudo de caso para dobramento de proteínas

Rio Grande - RS

23/05/2014

Renan Martinez da Luz

Descoberta de Conhecimento com Auxílio da Inteligência Humana: um estudo de caso para dobramento de proteínas

Universidade Federal do Rio Grande – FURG

Mestrado em Modelagem Computacional

Programa de Pós-Graduação

Orientador: Adriano Velasque Werhli

Coorientador: Diana Francisca Adamatti

Rio Grande - RS

23/05/2014

RENAN MARTINEZ DA LUZ

"DESCOBERTA DE CONHECIMENTO COM AUXÍLIO DA INTELIGÊNCIA HUMANA: UM ESTUDO DE CASO PARA DOBRAMENTO DE PROTEÍNAS"

Dissertação apresentada ao Programa de Pós Graduação em Modelagem Computacional da Universidade Federal do Rio Grande -FURG, como requisito parcial para obtenção do Grau de Mestre. Área concentração: Modelagem Computacional.

Aprovada em

BANCA EXAMINADORA

Prof. Dr. Adriano Velasque Werhli
Orientador - FURG

Prof.ª Dra. Karina dos Santos Machado
FURG

Prof.ª Dra. Regina Barwaldt
FURG

Prof.ª Dra. Carine Geltrudes Webber
UCS

Rio Grande -RS
2014

Agradecimentos

Agradeço aos meus familiares e pessoas próximas que me deram apoio para conseguir mais esta conquista.

Resumo

As proteínas desempenham um papel fundamental na natureza e a descoberta de suas funcionalidades e seus comportamentos ainda inexplorados, despertam muito o interesse de diversas áreas. Em 2010, o jogo sério chamado *"Fold it"* foi desenvolvido com objetivo de capturar técnicas de dobramento de proteína através da inteligência humana chegando a resultados surpreendentes.

Este trabalho tem como objetivo principal a busca por conhecimentos e técnicas de dobramento de proteínas através da inteligência humana, utilizando estruturas de proteínas no modelo HP possibilitando a predição dessas estruturas de forma simplificada, mas próxima a da realidade.

Para tanto foi elaborado um jogo sério de dobramento de proteína no modelo HP, onde foi possível obter dobramentos realizados pelos jogadores. Estes dados foram submetidos a técnicas de mineração de dados onde foi possível extrair conhecimentos e algumas estratégias de dobramento de proteínas.

Através dos resultados obtidos foi possível analisar que, é possível obter conhecimentos com a ajuda da inteligência humana, isso indica que no futuro existem chances de se desenvolver um jogador artificial que tenha o potencial de executar dobramentos melhores dos que foram adquiridos, com a ajuda de algoritmos.

Palavras-chaves: Dobramento de Proteínas. Jogo Sério. Mineração de dados. Inteligência Humana.

Abstract

Proteins are fundamental units in nature and the discovery of its unexplored functionalities and behaviors captures the attention of many areas of study. In 2010 the game called "fold it" was developed with the aim of capturing folding techniques using human intelligence. The results of this work were very promising.

The main aim of this work is to search for knowledge and techniques of protein folding using human intelligence. The HP protein model is used as a means to simplify the problem while keeping its main characteristics.

Following the main aim a serious game was developed where players have to fold an HP model protein. Using this game it was possible to record in a database the folding made by the players. This data set was analyzed with data mining techniques where some strategies and knowledge about the protein folding were extracted.

Analyzing the results we can observe that it is possible to obtain knowledge with the help from human intelligence. The results also indicates that further research has the potential to produce an artificial player that will have the ability to fold proteins better than traditional algorithms.

Key-words: Protein folding. Serious Game. Data Mining. Human Intelligence.

Lista de Figuras

Figura 1 – Fluxograma da metodologia aplicada neste projeto. 13
Figura 2 – Ligação peptídica 15
Figura 3 – Processo de síntese de proteínas (BRANDEN; TOOZE, 1999) 16
Figura 4 – Estrutura primária ou sequância linear de aminoácidos de uma proteína (BRANDEN; TOOZE, 1999) 17
Figura 5 – Estrutura Secundária de uma proteína contendo hélice-alfa e folha-beta (BRANDEN; TOOZE, 1999) 18
Figura 6 – Exemplos de estruturas terciarias (BRANDEN; TOOZE, 1999) 19
Figura 7 – Estrutura 3D de uma proteína (BRANDEN; TOOZE, 1999) 20
Figura 8 – Estrutura de uma proteína no modelo HP (DILL, 1985) 21
Figura 9 – Dobramento de uma estrutura 2D modelo HP (DILL, 1985) 22
Figura 10 – Ligação hidrofóbica em uma estrutura 2D modelo HP 23
Figura 11 – *Foldit* – jogo 3D de dobramento de proteína (BAKER, 2000) 25
Figura 12 – *Galaxy Zoo* – jogo com objetivo de localizar objetos celestes (BAKER, 2000) 25
Figura 13 – Processo KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) 26
Figura 14 – Fluxograma do projeto *Foldit* (COOPER et al., 2010b) 31
Figura 15 – Ambiente do jogo *Foldit* (COOPER et al., 2010b) 32
Figura 16 – Ambiente do jogo Eyewire (GARCIA, 2013) 33
Figura 17 – Como cientistas fotografam o cérebro (GARCIA, 2013) 34
Figura 18 – Capturas de telas do jogo Calangos em dois períodos simulados do dia, com a luz do sol e a noite. (LOULA et al., 2011) 35
Figura 19 – Fluxograma com as ferramentas utilizadas 37
Figura 20 – Modelo de relacionamento do banco de dados 38
Figura 21 – Interface do jogo 40
Figura 22 – Interface - Usuário logado 41
Figura 23 – Lista das proteínas que podem ser selecionadas para jogar 42
Figura 24 – Interface - jogo iniciado 43
Figura 25 – Exemplos do uso das ferramentas 44
Figura 26 – Interface - Ambiente do jogo com a estrutura dobrada 45
Figura 27 – Exemplo de um arquivo CSV 46
Figura 28 – Interface do software WEKA 47
Figura 29 – Exemplo das posições dos aminoácidos na estrutura da proteína modelo HP 53
Figura 30 – Árvore de classificação gerada no teste 1 54
Figura 31 – Árvore de classificação gerada no teste 2 versão 1 56
Figura 32 – Árvore de classificação gerada no teste 2 versão 2 58
Figura 33 – Árvore de classificação gerada no teste 3 versão 1 60

Figura 34 – Árvore de classificação gerada no teste 3 versão 2 62
Figura 35 – Árvore de classificação gerada no teste 4 versão 1 65
Figura 36 – Árvore de classificação gerada no teste 4 versão 2 67
Figura 37 – Árvore de classificação gerada no teste 5 versão 1 69
Figura 38 – Árvore de classificação gerada no teste 5 versão 2 71

Lista de tabelas

Tabela 1 – Os vinte tipos de aminoácidos (LEHNINGER; NELSON; COX, 2008) . . . 16
Tabela 2 – Sequências HP compiladas (HART; ISTRAIL, 2012) 22
Tabela 3 – Nº de Instâncias e Nº mínimo de instâncias por folha 48
Tabela 4 – Percentuais de instâncias corretamente classificadas 52
Tabela 5 – Matriz de confusão teste 1 53
Tabela 6 – Matriz de confusão teste 2 versão 1 55
Tabela 7 – Matriz de confusão teste 2 versão 2 57
Tabela 8 – Matriz de confusão teste 3 versão 1 59
Tabela 9 – Matriz de confusão teste 3 versão 2 61
Tabela 10 – Matriz de confusão teste 4 versão 1 63
Tabela 11 – Matriz de confusão teste 4 versão 2 66
Tabela 12 – Matriz de confusão teste 5 versão 1 68
Tabela 13 – Matriz de confusão teste 5 versão 2 70

Lista de abreviaturas

CSS	Cascading Style Sheets
DNA	Deoxyribonucleic Acid
HIV	Vírus da imunodeficiência humana
HP	Hidrofóbico e Polar
HTML	Hyper Text Markup Language
KDD	Knowledge Discovery in Databases ou Descoberta de Conhecimento nas Bases de Dados
MySql	My Structured Query Language
PHP	Personal Home Page
RNA	Ribonucleic Acid
SGBD	Sistema de Gerenciamento de Banco de Dados

Sumário

1 INTRODUÇÃO 12
1.1 Objetivo 13
1.2 Metodologia 13
1.3 Estrutura do Texto 14

2 REVISÃO BIBLIOGRÁFICA 15
2.1 Proteínas 15
2.2 Dobramento de Proteínas 19
2.3 Modelo Hidrofóbico-Polar Bidimensional 20
2.4 Jogos 23
2.4.1 Jogos Sérios 24
2.5 Mineração de Dados 26
2.5.1 Classificação 28
2.5.2 Associação 29
2.5.3 Agrupamento 29
2.6 Trabalhos relacionados 30
2.6.1 Foldit 30
2.6.2 Eyewire 32
2.6.3 Calangos 34
2.6.4 Mineração de dados na classificação e seleção de Oncogenes 35

3 MATERIAIS E MÉTODOS 37
3.1 Implementação de um jogo sério biológico 37
3.1.1 Implementação 37
3.1.2 Banco de dados 38
3.1.3 Interface 39
3.2 Aquisição dos Dados 45
3.3 Pré-processamento, transformação e mineração dos dados 45
3.3.1 Weka e Algoritmo J48 46
3.3.2 Características dos testes 48
3.3.3 Teste 1 49
3.3.4 Testes 2 49
3.3.5 Testes 3 49
3.3.6 Testes 4 50
3.3.7 Testes 5 50

4 RESULTADOS 52
4.1 Percentuais de instâncias corretamente classificadas 52
4.2 Teste 1 53
4.3 Testes 2 55
4.4 Testes 3 59
4.5 Testes 4 63
4.6 Testes 5 68
4.7 Considerações finais 72

5 CONCLUSÕES E TRABALHOS FUTUROS 73

Referências 74

1 Introdução

Atualmente, ao se tratar do problema de dobramento de proteínas, existe uma vasta gama de desafios ainda não resolvidos (COOPER et al., 2010b).

As proteínas desempenham um papel fundamental na natureza e essas estruturas, compostas de aminoácidos, participam em muitas tarefas importantes como garantir o correto funcionamento das células. A descoberta de suas funcionalidades e seus comportamentos ainda inexplorados, desperta muito o interesse de áreas envolvidas com a biologia, empresas de fabricação de medicamentos e com a produção de outras proteínas (PTITSYN, 1996).

A computação também está cada vez mais presente na solução destes problemas e esta tendência visa transformar estes problemas biológicos em modelos computacionais através de algoritmos, alcançando assim respostas relevantes do desconhecido. Mas, muitos destes algoritmos ainda não conseguem ser eficazes em seus resultados, devido a limitações tecnológicas, tanto físicas quanto lógicas (COOPER et al., 2010a).

Mediante as limitações tecnológicas para simular dobramentos das estruturas das proteínas, foi criado em 2010 o jogo sério chamado *"Fold it"* com objetivo de capturar técnicas de dobramento de proteína através da inteligência humana (COOPER et al., 2010a). A inteligência humana, em um ambito científico, é a capacidade de seres humanos de raciocinar, planejar, solucionar problemas e abstrair idéias por conta própria (MIRANDA, 2002). Sem mesmo saber a real solução do problema, as pessoas conseguem elaborar estratégias para chegar a algum resultado.

Pesquisadores da Universidade de Washington decidiram desbravar estratégias usadas pelos jogadores, e em setembro de 2011 desvendaram a estrutura de uma proteína que existe nos retrovírus, como o HIV. O papel dela é realizar a multiplicação dos vírus. Ao ser estudada, poderá ajudar a criar vacinas e remédios que evitem que os vírus atinjam células saudáveis.

Segundo (COOPER et al., 2010a): "Nós concedemos aos participantes a possibilidade de criar e melhorar as fórmulas para jogar o *Fold it*. Assim que vimos a variedade e a criatividade dessas fórmulas, usando até algoritmos, ficamos chocados. Para nós, isso é ainda mais emocionante do que a descoberta de setembro".

Devido as dificuldades de estudar as complexas estruturas das proteínas com todos os seus aminoácidos, foi criado por (DILL, 1985) o modelo HP (Hidrofóbico-Polar) que simplifica as estruturas em um modelo 2D que apresenta apenas uma sequência binária de aminoácidos com isso facilitando a predição das estruturas.

Portanto, devido à importância do assunto abordado e resultados positivos do uso da inteligência humana para buscar soluções de dobramento de proteína, este trabalho visa utilizar

a inteligência humana para estudar dobramentos de proteínas utilizando modelo HP, através de um jogo sério, para descobrir novos conhecimentos extraidos estatisticamente com a ajuda técnica de Árvore de decisão da mineração de dados, baseado nas estratégias dos jogadores, de forma a complementar os algoritmos já existentes.

1.1 Objetivo

O objetivo principal deste trabalho é a busca por conhecimentos e técnicas de dobramento de proteína através da inteligência humana utilizando estruturas no formato do modelo HP (Hidrofóbico-Polar).

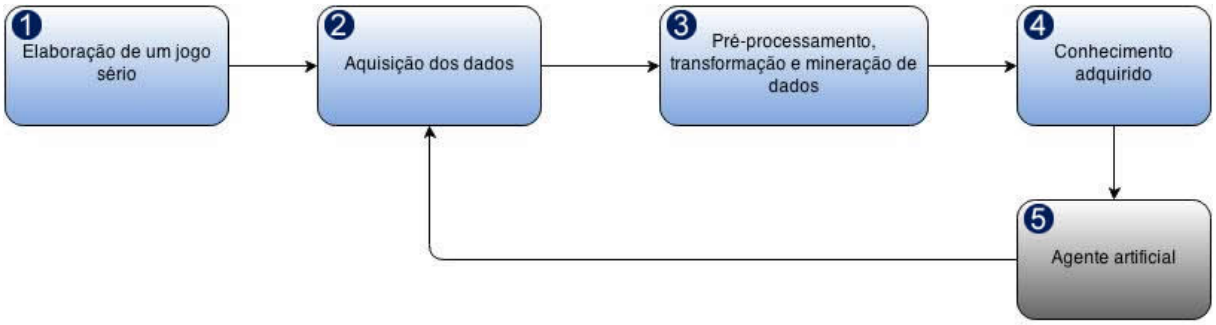
Para tanto, tem-se os seguintes objetivos específicos:

- Implementar um jogo sério biológico;
- Testar o jogo e armazenar as jogadas em um banco de dados;
- Selecionar e aplicar técnicas de mineração de dados sobre as instâncias (jogadas) adquiridas;
- Analisar os resultados obtidos;

1.2 Metodologia

A metodologia aplicada pode ser analisada resumidamente através do fluxograma (Figura 1), seguido de uma breve explicação de cada etapa seguida.

Figura 1: Fluxograma da metodologia aplicada neste projeto.



1. No primeiro passo foi desenvolvido um jogo sério biológico de dobramento de proteínas, onde estão disponíveis as estruturas de 20 proteínas diferentes, com suas sequências HP e suas energias ótimas que já foram determinadas por outros algoritmos, e que se encontram com suas posições iniciais (esticadas) possibilitando os jogadores executar uma série de dobramentos na estrutura.

2. Após todos os testes a nível computacional no jogo, este foi aplicado para alunos da graduação, com intuito de obter um volume significativo de entradas no banco de dados.

3. Após a obtenção de um bom volume de entradas no banco de dados, foi feito um processamento desses dados que foram submetidos a técnicas de mineração de dados, onde foi possível extrair alguns conhecimentos de possíveis técnicas de dobramento de proteína.

4. Em alguns casos foram adquirido apenas conhecimentos para melhorar a seleção de dados e submetendo-os novamente nas técnicas de mineração de dados. Nesta etapa, os resultados obtivos foram analisados com suas árvores de decisão, o percentual de acerto classificatório e as matrizes de confusão.

5. Esta etapa não faz parte do escopo do trabalho, mas o próximo passo deste projeto é desenvolver agentes artificiais, baseado no conhecimento adquirido.

1.3 Estrutura do Texto

Este texto está estruturado em 5 capítulos:

1. **Introdução:** Introduz o leitor, explicando e motivando aos assuntos abordados.

2. **Revisão Bibliográfica:** Contém algumas explicações a respeito do embasamentos teórico deste projeto com os seguintes assuntos: Proteínas, Dobramento de proteínas, Modelo HP, Jogos, Jogos Sérios, Mineração de dados. E também trabalhos relacionados com o assunto.

3. **Materiais e Métodos:** Contém os materiais e os métodos utilizados para a obtenção dos objetivos propostos.

4. **Resultados:** Apresentação, explicação e considerações finais dos resultados obtidos.

5. **Conclusões e Trabalhos Futuros:** Conclusão geral do projeto e trabalhos futuros.

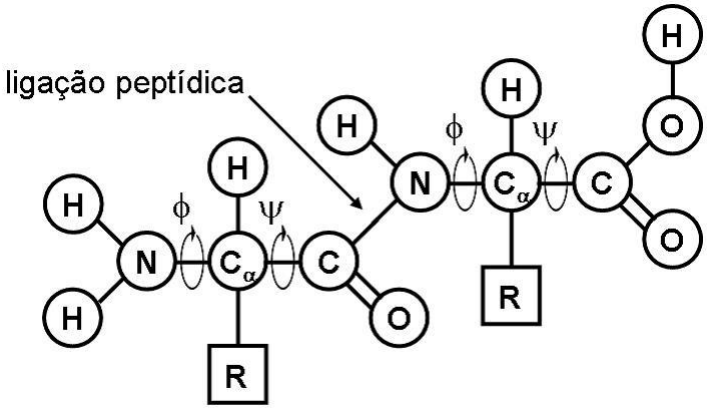
2 Revisão Bibliográfica

2.1 Proteínas

As proteínas são moléculas orgânicas abundantes e essenciais nas células sendo encontradas em todas as partes delas. São formadas pela associação de uma série determinada de aminoácidos, unidos entre si por ligações peptídicas.

Uma ligação peptídica é a união do grupo amino (-NH2) de um aminoácido com o grupo carboxila (-COOH) de outro aminoácido, através da formação de uma amida como podemos ver na Figura 2.

Figura 2: Ligação peptídica



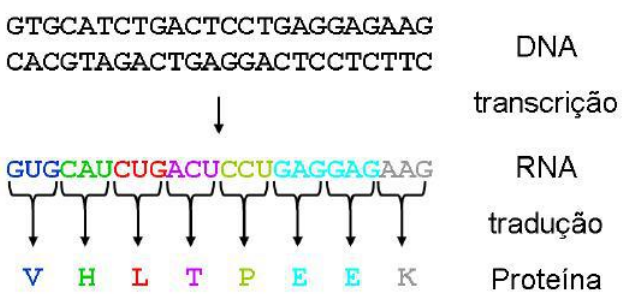
Na Tabela 1 pode-se ver os vinte tipos de aminoácidos existentes na natureza onde a maioria das proteínas é constituída por mais de 100 aminoácidos, sendo algumas delas constituídas por mais de 4000 (LEHNINGER; NELSON; COX, 2008). Pode-se observar também que cada aminoácido possui uma cadeia lateral diferente e por consequência, diferentes características físico-químicas. Uma das características mais importantes é a hidrofobicidade. Sendo assim, um aminoácido pode ser hidrofóbico ou hidrofílico. Os aminoácidos hidrofóbicos no dobramento tendem a se proteger da água buscando o núcleo da proteína, enquanto que os hidrofílicos tendem a ir para a superfície da proteína (BRANDEN; TOOZE, 1999; TROVATO et al., 2005).

Tabela 1: Os vinte tipos de aminoácidos (LEHNINGER; NELSON; COX, 2008)

Aminoácido	Abreviação	Código	Hidrofobicidade	Cadeia lateral
Alanina	Ala	A	Hidrofóbico	0
Cisteina	Cys	C	Hidrofílico	$-CH_2SH$
Ácido aspártico	Asp	D	Hidrofílico	$-CH_2COOH$
Ácido glutâmico	Glu	E	Hidrofílico	$-CH_2CH_2COOH$
Fenilalanina	Phe	F	Hidrofóbico	$-CH_2C_6H_5$
Glicina	Gly	G	Hidrofóbico	$-H$
Histidina	His	H	Hidrofílico	$-CH-C_3H_3N_2$
Isoleucina	Ile	I	Hidrofóbico	$-CH(CH_3)CH_2CH_3$
Lisina	Lys	K	Hidrofílico	$-(CH_2)_4NH_2$
Leucina	Leu	L	Hidrofóbico	$-CH_2CH(CH_3)_2$
Metionina	Met	M	Hidrofóbico	$-CH_2CH_2SCH_3$
Asparagina	Asn	N	Hidrofílico	$-CH_2CONH_2$
Prolina	Pro	P	Hidrofóbico	$-CH_2-CH_2-CH_2$
Glutamina	Gln	Q	Hidrofílico	$-CH_2CH_2CONH_2$
Arginina	Arg	R	Hidrofílico	$-(CH_2)_3NH-C(NH)NH_2$
Serina	Ser	S	Hidrofílico	$-CH_2OH$
Treonina	Thr	T	Hidrofílico	$-CH(OH)CH_3$
Valina	Val	V	Hidrofóbico	$-CH(CH_3)_2$
Triptofano	Trp	W	Hidrofóbico	$-CH_2C8H_6N$
Tirosina	Tyr	Y	Hidrofóbico	$-CH_2-C_6H_4OH$

De acordo com (LEHNINGER; NELSON; COX, 2008), os aminoácidos são gerados dentro das células pelo processo de decodificação do RNA, chamado de síntese de proteínas que é um processo rápido, que ocorre em todas as células do organismo, mais precisamente, nos ribossomos, organelas encontradas no citoplasma e no retículo endoplasmático rugoso sendo dividido em dois processos mostrados na Figura 3. Estes processos serão discutidos a serguir.

Figura 3: Processo de síntese de proteínas (BRANDEN; TOOZE, 1999)



A) Transcrição

A mensagem contida no cístron (porção do DNA que contém a informação genética necessária à síntese proteica) é transcrita pelo RNA mensageiro (RNAm). Nesse processo, as bases pareiam-se: a adenina do DNA se liga à uracila do RNA, a timina do DNA com a adenina do RNA, a citosina do DNA com a guanina do RNA, e assim sucessivamente, havendo a intervenção da enzima RNA-polimerase. A sequência de 3 bases nitrogenadas de RNAm, forma o códon, responsável pela codificação dos aminoácidos. Dessa forma, a molécula de RNAm replica a mensagem do DNA, migra do núcleo para os ribossomos, atravessando os poros da membrana plasmática e forma um molde para a síntese proteica.

B) Tradução

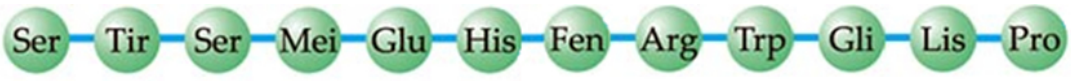
Na fase de tradução, a mensagem contida no RNAm é decodificada e o ribossomo a utiliza para sintetizar a proteína de acordo com a informação dada. Os ribossomos são formados por duas subunidades. Na subunidade menor, ele faz ligação ao RNAm, na subunidade maior há dois sítios (1 e 2), em que cada um desses sítios podem se unir a duas moléculas de RNAt. Uma enzima presente na subunidade maior realiza a ligação peptídica entre os aminoácidos, o RNA transportador volta ao citoplasma para se unir a outro aminoácido. E assim, o ribossomo vai percorrendo o RNAm e provocando a ligação entre os aminoácidos.

O fim do processo ocorre quando o ribossomo passa por um códon de terminação e nenhum RNAt entra no ribossomo, por não terem mais sequências complementares aos códons de terminação. Então, o ribossomo se solta do RNAm, a proteína específica é formada e liberada do ribossomo.

As proteínas são moléculas bastante complexas exibindo uma hierarquia de níveis de organização estrutural. Os níveis de organização estruturais são:

- Estrutura primária: sequência de aminoácidos que compõe a cadeia polipeptídica conforme mostra a Figura 4.

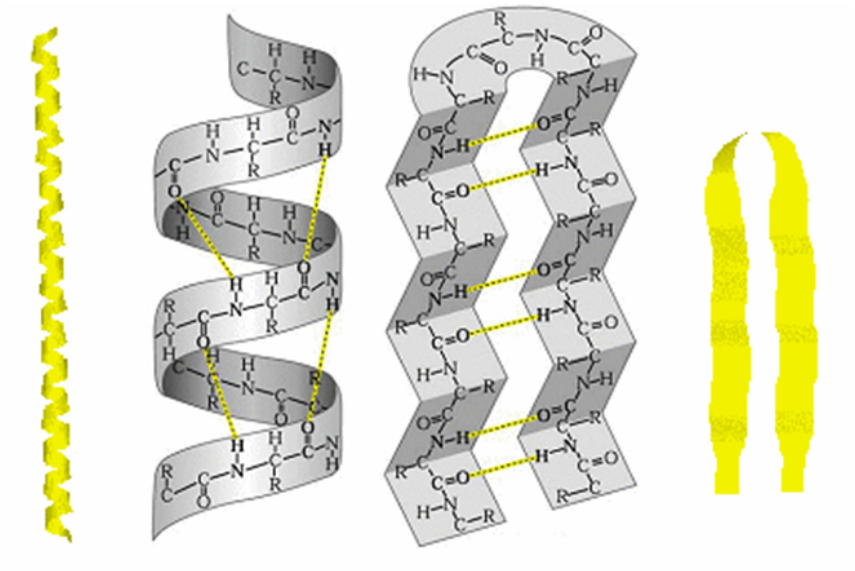
Figura 4: Estrutura primária ou sequância linear de aminoácidos de uma proteína (BRANDEN; TOOZE, 1999)



- Estrutura secundária: onde a cadeia polipeptídica é flexível podendo rodar em torno de três ligações químicas sobre os quais, a proteína pode rodar acedendo a diferentes configurações espaciais.

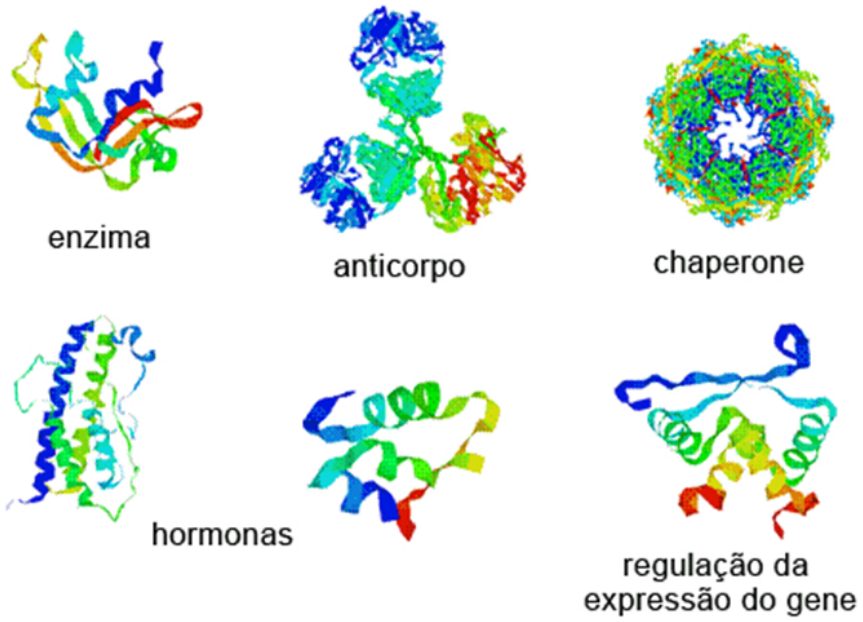
De todas as configurações dimensionais que as proteínas podem apresentar as mais comuns são as alfa-hélice e a folha-beta como vê-se na Figura 5, que para além das ligações peptídicas, também envolve as chamadas ligações por pontes de hidrogénio, um tipo de interação eletrostática entre os átomos de hidrogénio e oxigênio, chama-se elementos de estrutura secundária.

Figura 5: Estrutura Secundária de uma proteína contendo hélice-alfa e folha-beta (BRANDEN; TOOZE, 1999)



- E estrutura terciária da proteína: resultado das interações hidrofóbicas entre os aminoácidos e a água é a junção e o empacotamento dos elementos de estrutura secundária. A disposição tridimensional dos átomos da proteína na estrutura terciária é de extrema importância porque geralmente coincide com a chamada estrutura nativa, a estrutura que concede à proteína uma função biológica específica. Pode-se ver alguns exemplos dessas estruturas na Figura 6.

Figura 6: Exemplos de estruturas terciarias (BRANDEN; TOOZE, 1999)

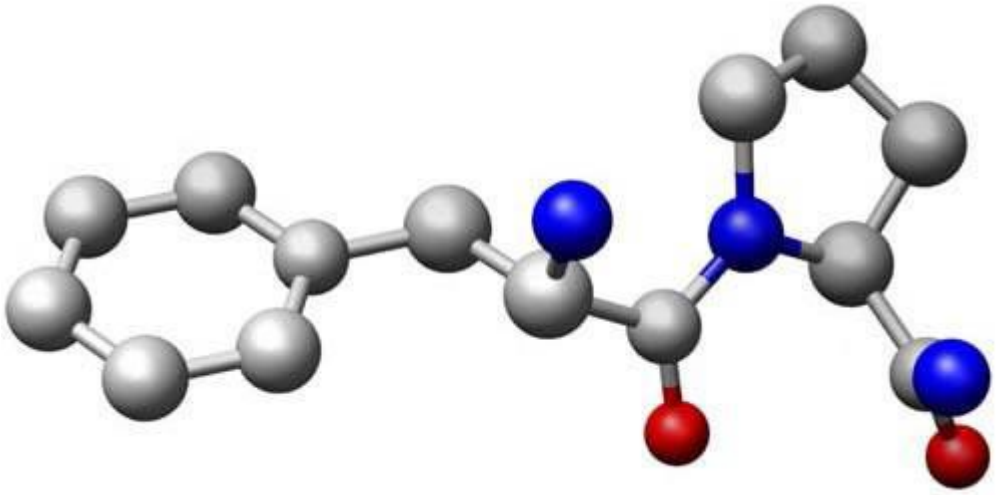


2.2 Dobramento de Proteínas

O dobramento de proteína é um processo onde uma proteína assume a sua configuração funcional através de sua estrutura.

As moléculas de proteínas são cadeias heterogêneas não ramificadas de aminoácidos e para que possa desempenhar uma função, a estrutura primária deve assumir forma tridimensional específica. As proteínas executam o processo chamado de dobramento obtendo um aspecto como o ilustrado na Figura 7. Sendo assim são capazes de realizar a sua função biológica que podem servir para a criação de grandes módulos de montagens, tais como, partículas de vírus ou de fibras musculares, ou podem proporcionar locais específicos de ligação, como os encontrados em enzimas ou proteínas que transportam oxigênio e até mesmo que regulam a função de DNA (BRANDEN; TOOZE, 1999; TROVATO et al., 2005).

Figura 7: Estrutura 3D de uma proteína (BRANDEN; TOOZE, 1999)



Neste processo de dobramento de proteína, o comportamento dos aminoácidos depende de certas soluções de acordo com o ambiente em que se encontra, incluindo o tipo de solvente primário no interior das células que pode ser água ou lipídeos, da concentração dos sais, da temperatura e das moléculas que a rodeiam (BAKER, 2000; CHANDRU; DATTASHARMA; KUMAR, 2003).

Acredita-se que o processo de dobramento de proteína não contém resultados aleatórios e que apresenta um complexo mecanismo natural com conformações nativa mais estável e com um número menor de energia livre possível. E também mostra que o processo de dobramento deve ser realizado em apenas um período de tempo utilizando uma quantidade finita de movimentos (DINNER et al., 2000; PLOTKIN; ONUCHIC, 2000).

Portanto, a compreensão destas características é de suma importância para varias áreas como a da saúde, onde seria possível a criação de drogas inteligentes, entendimento de algumas doenças como câncer, mal de Alzheimer, mal de Parkinson e diabetes tipo II, que são causadas por proteínas aglomeradas e dobradas incorretamente não desempenhando corretamente sua função em nosso organismo (BAKER, 2000; CHANDRU; DATTASHARMA; KUMAR, 2003; DINNER et al., 2000; PLOTKIN; ONUCHIC, 2000; TROVATO et al., 2005).

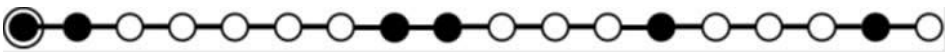
2.3 Modelo Hidrofóbico-Polar Bidimensional

O Modelo Hidrofóbico-Polar Bidimensional (Modelo HP) é um modelo que tem como característica trabalhar com uma sequência binária de aminoácidos "H"(hidrofóbicos, apolares) ou "P"(hidrofílicos, polares) reduzindo o alfabeto de vinte aminoácidos em apenas dois. Baseia-se na crença de que a maior contribuição para a energia da conformação nativa de uma proteína é devido às interações entre os aminoácidos hidrofóbicos, que tendem a se proteger de algum

solvente de seu ambiente, movendo-se para o núcleo da estrutura e sendo envolvidos pelos aminoácidos hidrofílicos que tendem a permanecer na superfície da estrutura 3D (DILL, 1985; DILL et al., 1995).

Na Figura 8 pode-se analisar uma estrutura simplificada seguindo o Modelo HP, onde os resíduos são representados por círculos. Os círculos pretos representam os aminoácidos "H"e os círculos brancos representam os aminoácidos "P". Devido à teoria do modelo proposto, a estrutura encontra-se sobre os nós de uma grade onde os dois primeiros resíduos são fixos e os demais resíduos podem ser dobrados em ângulos de 90°, mas sempre mantendo o tamanho das ligações dos aminoácidos que são representados por traços pretos entre os circulos obedecendo a extensão de um quadrante da grade.

Figura 8: Estrutura de uma proteína no modelo HP (DILL, 1985)



Apesar da simplicidade do modelo HP, o processo de dobramento têm semelhanças de comportamento com o processo de dobramento no sistema real. O modelo HP tem sido usado pelos químicos para avaliar novas hipóteses de formação de estrutura das proteínas (DILL, 1985; DILL et al., 1995).

O Benchmark (HART; ISTRAIL, 2012) é uma compilação de sequências onde é possível comparar resultados de vários modelos discretos ajudando a chegar na solução de problemas ainda não resolvidos a respeito das estruturas das proteínas. Na Tabela 2 pode-se ver a quantidade de resíduos e a menor energia encontrada para cada sequência.

Tabela 2: Sequências HP compiladas (HART; ISTRAIL, 2012)

Nome	Sequencia	Tamanho	Energia
S1	$H_2P_5H_2P_3HP_3HP$	18	-4
S1	$HPHPH_3P_3H_4P_2H_2$	18	-8
S3	$PHP_2HPH_3PH_2PH_5$	18	-9
S4	$HPHP_2H_2PHP_2HPH2P_2HPH$	20	-9
S5	$H_3P_2HPHPHP_2HPHPHPPH$	20	-10
S6	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	24	-9
S7	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	25	-8
S8	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	36	-14
S9	$P_2HP_2H_2P_2H_2P_5H_10P_6H_2P_2H_2P_2HP_2H_5$	48	-23
S10	$H_2PHPHPHPH_4PHP_3HP_3HP_4HP_3HP_3HPH_4PHPHPHPH_2$	50	-21
S11	$P_2H_3PH_8P_3H_10PHP_3H_12P_4H_6PH_2PHP$	60	-36
S12	$H_12PHPHP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HPHPH_12$	64	-42
S13	$H_4P_4H_12P_6H_12P_3H_12P3H_12P_3HP_2H_2P_2H_2P_2HPH$	85	-53
S14	$P_6HPH_2P_5H_3PH_5PH_2P_4H_2P_2H_2PH_5PH_13PH_7P_11H_7P_2HPH_3P_6HPH_2$	100	-48
S15	$P_3H_2P_2H_4P_2H_3PH_2PH_2PH_4P_8H_6P_2H_6P_9HPH_2PH_11P_2H_3PH_2PHP_2HPH_3P_6H_3$	100	-50

A energia de uma estrutura no modelo HP depende do número e posição de aminoácidos hidrofóbicos que ocupam os pontos de uma grande onde a estrutura se encontra. A Figura 9 mostra um modelo HP 2D em que a estrutura da proteína ilustrada, contém 18 aminoácidos simplificados que ocupam os pontos da grade e 9 destes aminoácidos são hidrofóbicos. Se existem dois aminoácidos hidrofóbicos em pontos de grade adjacentes, mas que não são adjacentes na sequência de aminoácidos então se forma uma ligação hidrofóbica como indica a Figura 10 (DILL, 1985; DILL et al., 1995).

Figura 9: Dobramento de uma estrutura 2D modelo HP (DILL, 1985)

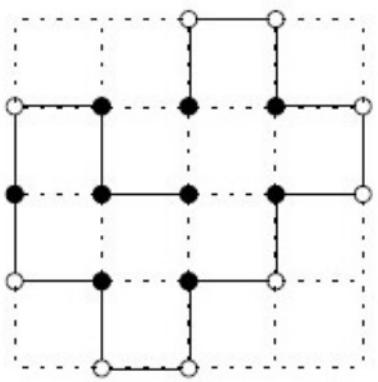
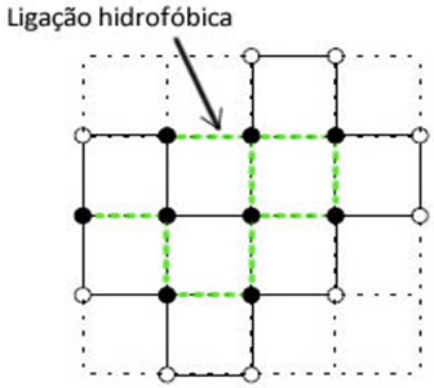


Figura 10: Ligação hidrofóbica em uma estrutura 2D modelo HP



A cada ligação hidrofóbica adquirida através de um dobramento no modelo HP, a estrutura da proteína atinge um valor melhor de energia. A cada movimento de dobramento, corre-se o riscos de acarretar perdas em outras ligações hidrofóbicas já existentes e também ao cruzamento da estrutura.

Portanto, é necessário a criação de técnicas de dobramento para obter um número ótimo de energia, priorizando uma maior compactação dos aminoácidos hidrofóbicos no interior da estrutura dobrada obtendo um maior número de ligações hidrofóbicas e evitando sua exposição ao solvente em seu ambiente (DILL, 1985; DILL et al., 1995).

2.4 Jogos

Segundo (HUIZINGA, 1971), o jogo faz parte da natureza do ser humano sendo essencial para o raciocínio, pois elementos lúdicos estão na base do surgimento e desenvolvimento da civilização. Ele define o jogo como: "uma atividade voluntária exercida dentro de certos e determinados limites de tempo e espaço, segundo regras livremente consentidas, mas absolutamente obrigatórias, dotado de um fim em si mesmo, acompanhado de um sentimento de tensão e alegria e de uma consciência de ser diferente de vida cotidiana.".

Cada vez mais a indústria tecnológica cresce devido o grande sucesso de jogos de entretenimento e aplicativos que estão sendo utilizados por toda a humanidade em seus computadores e aparelhos móveis. Com isso, aparecem oportunidades também para os jogos sérios despertarem o interesse dos jogadores (BAKER, 2000).

Os jogos estão sendo desenvolvidos com diversos objetivos, como diz Neusa Fialho que atua na área da educação: "O jogo exerce uma fascinação sobre as pessoas, que lutam pela vitória procurando entender os mecanismos dos mesmos, o que constitui de uma técnica onde os alunos aprendem brincando.". Mas não é apenas na área da educação. Empresas utilizam

jogos como ferramenta para buscar uma melhor performance de seus funcionários. Cientistas também utilizam os jogos, buscando respostas para seus problemas científicos (FIALHO, 2007).

Basicamente os jogos costumam ser estudados por quatro áreas do conhecimento humano: O antropológico, que estuda o significado e o contexto dos jogos; o sociológico, que estuda os efeitos dos jogos sobre as pessoas (aprendizado, desenvolvimento cognitivo, agressividade, etc); o tecnológico, que estuda os elementos que compõe os jogos e sua utilização analisando sua utilização como vetores de inovações tecnológicas e o comercial que analisa a criação, evolução e a comercialização dos jogos (ALLé, 1999).

2.4.1 Jogos Sérios

Assim como jogos desenvolvidos como técnicas de aprendizagem, jogos sérios ou *"Serious Games"* estão sendo desenvolvidos e utilizados no meio acadêmico com o objetivo de alcançar a resultados científicos, visando chegar a respostas de problemas ainda não resolvidos em diversas áreas de pesquisa, como a biologia (BAKER, 2000).

Um jogo sério é um *software* desenvolvido com o objetivo de transmitir um conteúdo de caráter educativo ao utilizador. O termo "Sério"refere-se neste caso a produtos e situações ligadas em áreas como a da educação, exploração científica, serviços de saúde, gestão de emergência, planejamento urbano, engenharia, religião, política e entre outras. O primeiro Jogo Sério foi o *Army Battlezone*, um projeto desenvolvido pela empresa Atari nos anos 80. Este jogo foi concebido para treinar militares em situação de batalha. Ao longo dos anos, e à medida que os computadores para uso pessoal foram desenvolvidos, os jogos sérios foram se espalhando para uma maior variedade de áreas: educação, treino profissional, saúde, publicidade, e políticas públicas (PRENSKY, 2004).

Um exemplo de Jogo sério é o jogo da equipe do *Foldit* Figura 11, afirma que jogos estão sendo usados para ajudar a ciência, aproveitando habilidades humanas de reconhecimento de imagens, por exemplo, para localizar objetos celestes, o *Galaxy Zoo* Figura 12. Este trabalho apresenta uma classe mais geral de descobertas científicas que se concentram em alavancar problemas humanos, resolvendo problemas científicos computacionalmente (COOPER et al., 2010b).

Figura 11: *Foldit* – jogo 3D de dobramento de proteína (BAKER, 2000)

Figura 12: *Galaxy Zoo* – jogo com objetivo de localizar objetos celestes (BAKER, 2000)

Um jogo de descoberta do campo da ciência transforma uma classe de problemas científicos difíceis em quebra-cabeças, e fornece em um mecanismo de jogo, para ajudar os jogadores não-especialistas a resolver estes problemas. Muitos aspectos tradicionais de *design* aplicam a jogos de descoberta científica, incluindo o design de níveis introdutórios para ajudar os recém-chegados e explicar a mecânica de jogo, o uso de uma arquitetura cliente-servidor para a competição e colaboração, e a exigência de que o jogo tem de ser divertido. No entanto, diferentemente dos jogos cujo objetivo é o entretenimento ou educação, jogos de descoberta científica introduzem um desafio único: permitir que não especialistas solucionadores de problemas naturais, obtenham um domínio científico específico (BAKER, 2000).

Portanto, chega-se a conclusão de que uma maneira de coletar dados, é elabora um jogo que através de uma interface intuitiva que utiliza um design que cultiva padrões comuns a muitos outros projetos já elaborados (NIELSEN, 2000), sendo assim trazendo o mecanismo do jogo em uma interface que facilite a jogabilidade do jogo, permitindo aos jogadores solucionar problemas científicos ainda não resolvidos.

2.5 Mineração de Dados

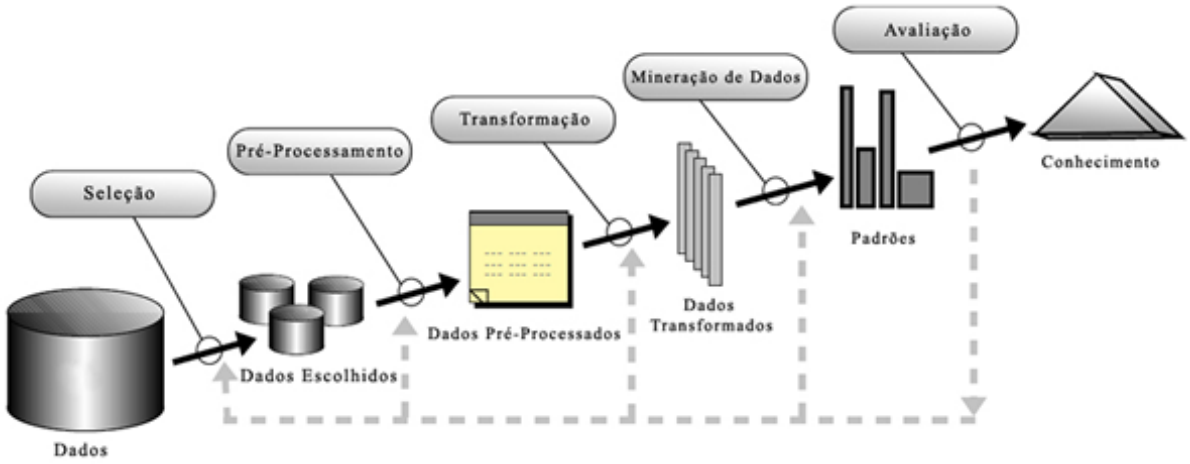
Conforme passa o tempo, a tecnologia avança em um nível acelerado, exigindo dos sistemas computacionais, um grau elevado de organização de dados devido aos seus números os volumes. Portanto, novas e mais complexas estruturas de armazenamento foram e estão sendo desenvolvidas, tais como: banco de dados, Data Warehouses, Bibliotecas Virtuais, Web e outras (CIOS et al., 2007; HAN; KAMBER, 2006).

Na década de 80, em busca de soluções, cientistas acharam algumas soluções e propuseram a criação da mineração de dados, termo vindo do inglês, *data mining*. E ao longo do tempo até os dias de hoje, a Mineração de Dados vem se mostrando promissora devido a grande demanda encontrada hoje em dia (LAROSE, 2005).

Muitas empresas já investiram em armazenamentos de dados devido a sua necessidade de obter resultados que os beneficiam, mas isso tudo sem sucesso. (HAN; KAMBER, 2006) identifica essa situação como "rico em dados, pobre em informação". Portanto foi elaborado por pesquisadores o processo KDD.

O KDD (Knowledge Discovery in Databases ou Descoberta de Conhecimento nas Bases de Dados) é uma tentativa de solucionar o problema de tratamento de grandes quantidades de dados. Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), é impraticável seguir os modelos tradicionais, onde para a transformação de conhecimento de um grande volume de dados, é necessário realizar um processo manual feito por cientistas com o objetivo de produzir relatórios em busca de conhecimento. Então, o KDD é uma tentativa de solucionar os problemas gerados por grandes volumes de dados.

Figura 13: Processo KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)



Uma das definições mais utilizadas para o termo KDD o define como "um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis"(FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A Figura 13 representa o processo KDD.

Alguns dos passos básicos de KDD serão explicados sucintamente de acordo com o autor já citado anteriormente:

1. Desenvolver um entendimento do domínio da aplicação e o conhecimento relevante já existente e identificar o objetivo do processo KDD do ponto de vista do cliente.
2. Criar um conjunto de dados alvo: selecionando um conjunto de dados ou focando em um subconjunto de variáveis ou amostras com as quais a descoberta será feita.
3. Limpeza e o processamento removendo ruídos, coletando informações necessárias para modelar ou levar em conta o ruído decidindo estratégias para lidar com dados não existentes.
4. Redução e projeção dos dados encontrando feições uteis para representar os dados dependendo do objetivo da tarefa.
5. Fazer a correspondência do objetivo do processo KDD com o método de mineração de dados em particular.
6. Análise exploratória e seleção de modelos e hipóteses escolhendo os algoritmos de mineração de dados e selecionando os métodos para procurar padrões de dados.
7. Mineração de dados procurando por padrões de interesse em uma forma representacional particular ou um conjunto de tais representações incluindo regras de classificação ou arvores regressão ou agrupamento.
8. Interpretação dos padrões minerados possivelmente retornando para um dos passos entre 1 e 7 para mais uma interação.
9. Agir com o conhecimento descoberto: utilizando-o diretamente, incorporando o conhecimento em outro sistema ou simplesmente documentando e repostando para interessados.

Segundo (CABENA et al., 1998), de uma perspectiva de banco de dados, a mineração de dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados.

Os principais objetivos das práticas de mineração de dados é a predição e a descrição. A predição envolve utilizar algumas variáveis ou campos do banco de dados para prever valores futuros ou desconhecidos de outras variáveis de interesse. A descrição foca em encontrar padrões que descrevem os dados e que sejam de passível de interpretação pelos humanos. Os objetivos da predição e descrição podem ser alcançados usando uma variedade de métodos de mineração de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Os diversos métodos de mineração de dados são classificados de formas diferentes dependendo do autor, aqui será usada a classificação adotada por (HAN; KAMBER, 2006; AGRAWAL; IMIELINSKI; SWAMI, 1993).

2.5.1 Classificação

A técnica de classificação consiste em aprender a função que mapeia/classifica um item em uma de várias classes pré-definidas. As técnicas de classificação podem ser supervisionadas e não supervisionadas e são usadas para prever valores de variáveis do tipo categóricas.

Alguns dos métodos de classificação mais utilizados são:

1. Classificação por árvore de decisão
   Funciona como um fluxograma em forma de árvore, onde cada nó indica um teste sobre o valor. Com o objetivo de reduzir a impureza ou incerteza dos dados o máximo possível.

   O J48, por exemplo, é uma implementação de código aberto do algoritmo de árvore de decisão C4.5. Esse algoritmo compreende as seguintes etapas.

   a) Verifica quais são os casos bases referente a cada coluna.

   b) Para cada atributo calcula-se o ganho de informação ao repartir a árvore por esse atributo. O ganho de informação é calculado pela diferença de entropia nos dados antes e após a repartição por um dado atributo. A entropia H do conjunto de dados D original é dada por:

$$H[D] = -\sum_{j=1}^{|C|} P(C_j) \log_2 P(C_j) \tag{2.1}$$

   E a entropia desse conjunto repartido por um dado atributo A é dada por:

$$H_A[D] = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} H[D_j] \tag{2.2}$$

   Onde v é o número de sub-conjuntos e C é o conjunto da classe desejada

   c) Escolhe-se o atributo que tem o maior ganho de informação

   d) Cria um nó de decisão que reparte a árvore utilizando o atributo de maior ganho

   e) Aplica-se os passos 2, 3 e 4 recursivamente para cada ramificação criada até que uma das seguintes características sejam encontradas:
   - Todos exemplos de um dado nó pertencem a mesma classe;
   - Não existem mais atributos para particionar;
   - Não existem mais exemplos;

2. Classificação por regressão
   A análise de regressão pode ser usada para modelar a relação entre uma ou mais variáveis independentes e uma variável dependente contínua. No contexto na mineração de dados, as variáveis independentes ou preditoras são os atributos de interesse, e em geral esses valores são conhecidos. A variável dependente é o que se quer prever.

3. Classificação baseada em regras
   A classificação baseada em regras segue a estrutura: SE condição ENTÃO conclusão semelhante as regras de associação. Esse tipo de construção geralmente é recuperado de uma árvore de decisão.

4. Classificação por regras de associação
   A ideia geral é buscar por padrões de associações fortes entre os itens, utilizando-se do conceito de frequência e as categorias.

2.5.2 Associação

O método de associação consiste em identificar o relacionamento dos itens mais frequentes em um determinado conjunto de dados. Um método de associação é a mineração de itens frequentes. Essa técnica pode ser visualizada em duas etapas: primeiro, um conjunto de itens frequentes (Frequent Itemset) é criado, respeitando um valor mínimo de frequência para os itens. Depois, as regras de associação são geradas pela mineração desse conjunto (HAN; KAMBER, 2006).

2.5.3 Agrupamento

O processo de agrupamento de um conjunto de objetos físicos ou abstratos em classes de objetos similares é chamado de *clustering* ou agrupamento. Um clustering é uma coleção de objetos similares uns aos outros dentro do mesmo cluster e dissimilares aos objetos em outros clusters. Alguns dos métodos de agrupamento mais utilizados de acordo com (HAN; KAMBER, 2006), são:

1. Métodos de particionamento
   Dado um banco de dados com N objetos, um método de particionamento constrói K partições dos dados onde cada partição representa um cluster e K menor ou igual N. Ou seja, os dados são classificados em k grupos que junto satisfazem os seguintes requisitos: 1) cada grupo deve conter apenas um objeto e 2) cada objeto deve pertencer a exatamente um grupo.

2. Métodos hierárquicos
   Os métodos hierárquicos criam uma decomposição hierárquica do conjunto de dados,

um método hierárquico pode ser classificado como aglomerativo ou divisivo. O aglomerativo também chamado de bottom up começa com cada objeto formando um grupo separado. Os objetos próximos uns aos outros são mesclados até que todos os grupos sejam mesclados em um único. O divisivo também chamado de top down começam com todos os dados no mesmo cluster em cada interação um cluster é dividido em clusters menores até que cada objeto esteja em um cluster diferente.

3. Método baseado em densidade
   A maioria dos métodos de particionamento agrupam os objetos baseado na distância entre eles, assim encontram clusters em formato esféricos com dificuldade para cluster de formas arbitrarias. Outros métodos de agrupamento foram desenvolvidos baseados na noção de densidade. A ideia geral é continuar crescendo um dado cluster até que a densidade na vizinhança exceda um limite.

2.6 Trabalhos relacionados

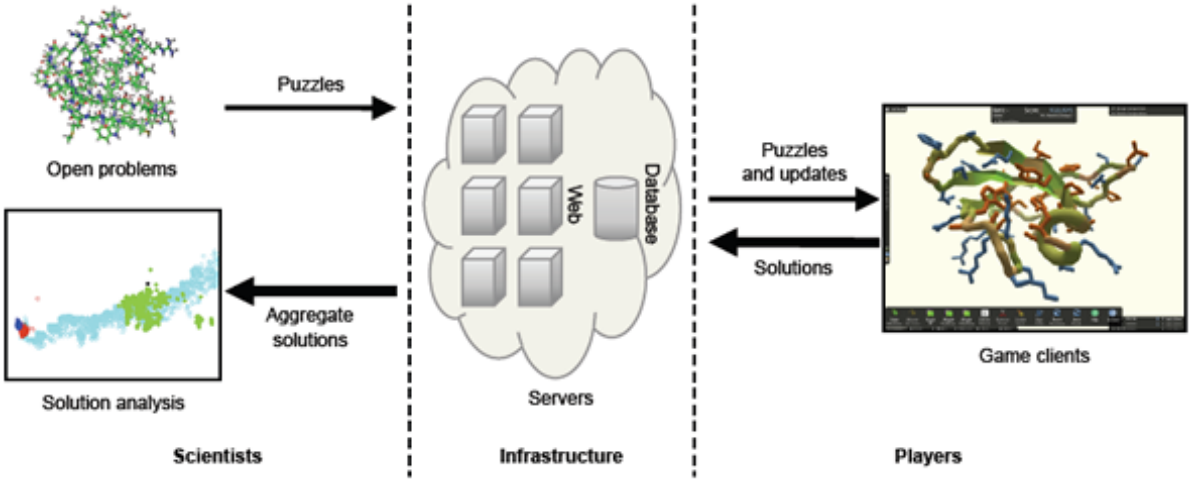
2.6.1 Foldit

O *Foldit* [1] é um projeto liderado por David Baker e Seth Cooper que tem como objetivo extrair técnicas de dobramento de proteínas com estrutura 3D através de um jogo sério.

A visão geral da arquitetura do projeto científico de Baker e Cooper está representada na Figura 14 que apresenta um fluxograma onde a equipe em seus trabalhos publicados explicam as finalidades do projeto, onde cientistas fornecem problemas de dobramento de proteína, disponibilizando-os para jogadores na Internet através de um jogo intuitivo, que por consequência os jogadores retornam possíveis soluções que são analisadas pelos cientistas (COOPER et al., 2010b).

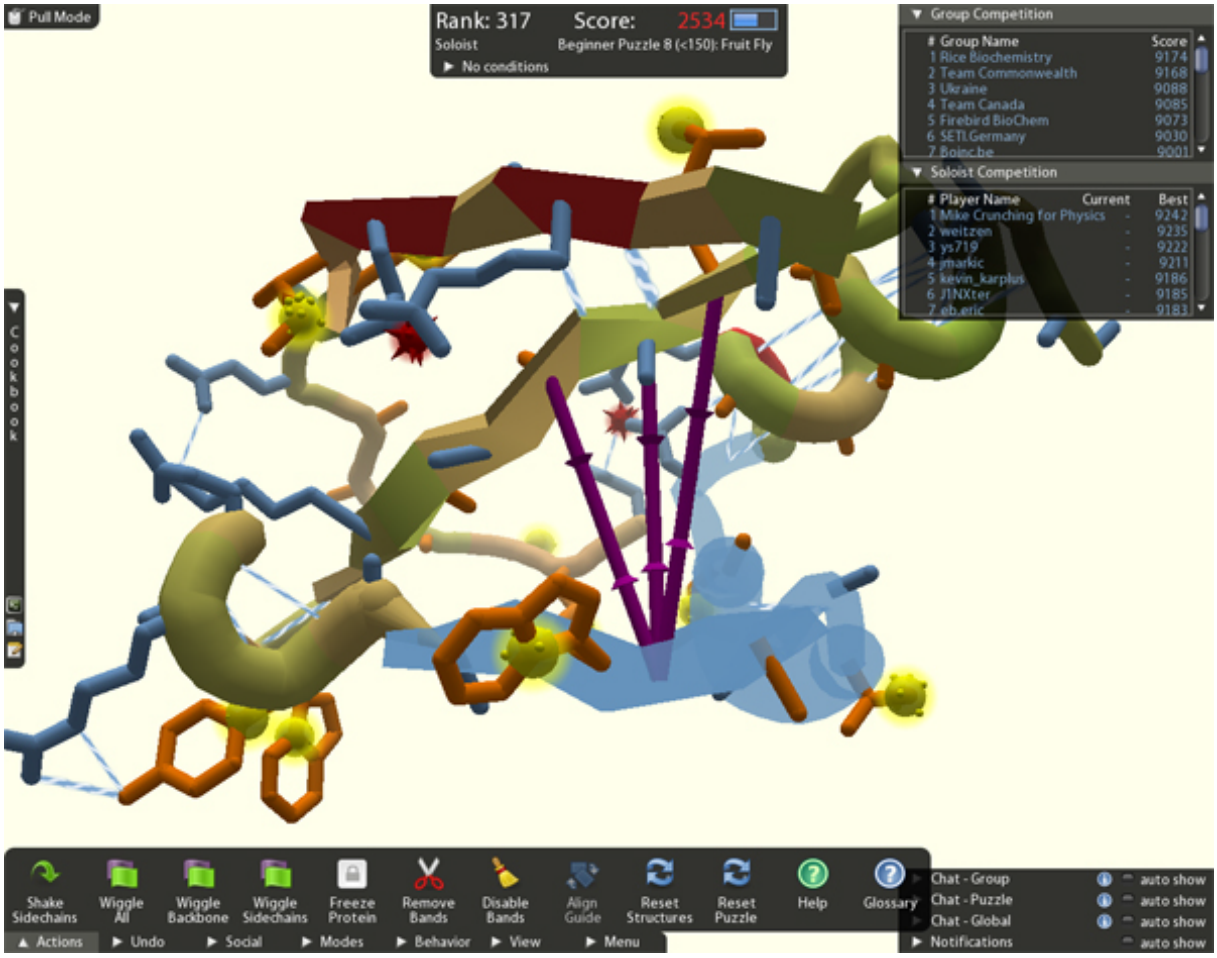
[1] https://fold.it/portal/

Figura 14: Fluxograma do projeto *Foldit* (COOPER et al., 2010b)



Mesmo sem nenhum conhecimento sobre o processo e sobre as áreas científicas envolvidas, as pessoas podem jogar e colaborar na descoberta de novas estruturas protéicas que possam ser utilizadas em benefício da humanidade. Além disso, como pode-se ver na Figura 15, onde mostra-se o ambiente do jogo, ele provem de muitas ferramentas de interface para jogadas e o compartilhamento de estratégias e de algoritmos entre os jogadores para ajudá-los a alcançar as melhores soluções para cada proteína e também fornece a mínima energia atual das proteínas por meio de um ranking tornando o jogo um pouco mais interessante e competitivo, uma atração para os jogadores (FOLDIT, 2012; COOPER et al., 2010b; KHATIB et al., 2011).

Figura 15: Ambiente do jogo *Foldit* (COOPER et al., 2010b)



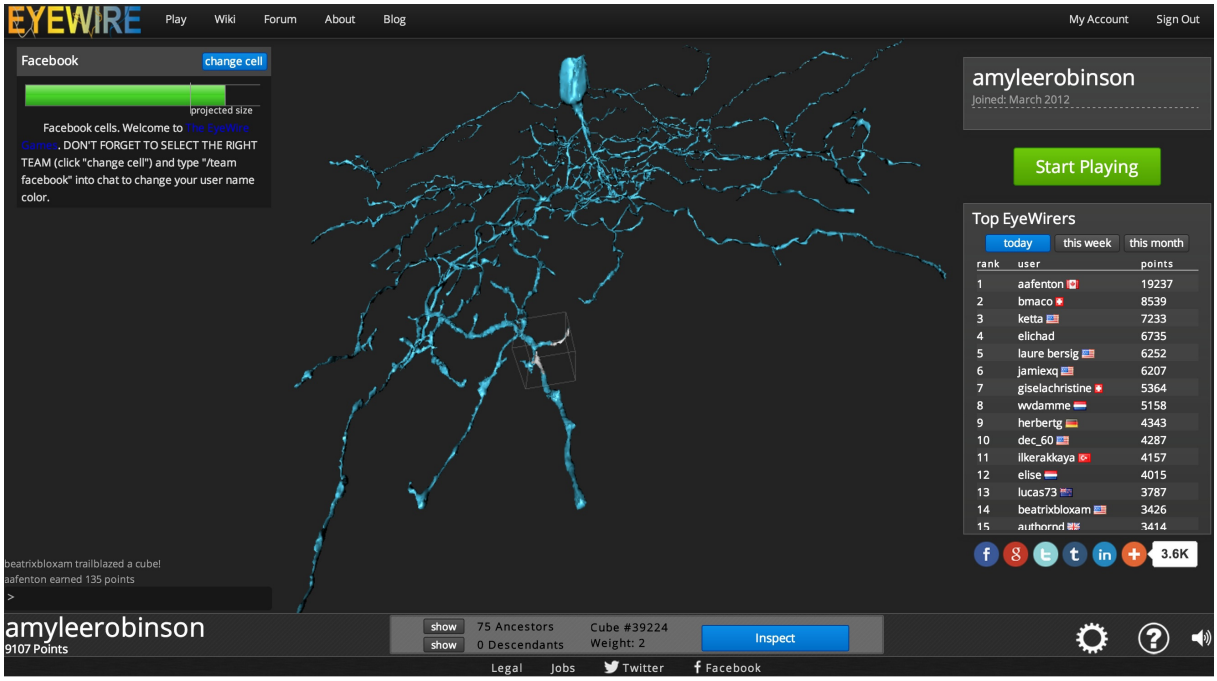
Em uma de suas conclusões, a equipe do *Foldit* informou que o projeto por ser diferente da maioria dos jogos, onde o entretenimento é o principal objetivo, o projeto *Foldit* foi focado principalmente para permitir que qualquer pessoa possa chegar à resolução de um problema científico. Ressaltam também que um dos aspectos mais difíceis de desenvolvimento pois estavam criando um jogo em que o resultado final não era conhecido.

2.6.2 Eyewire

O jogo Eyewire [2] (Figura 16), foi desenvolvido pelo laboratório do neurocientista Sebastian Seung do MIT, que tem como objetivo contribuir no mapeamento de conexões entre as células nervosas do organismo humano desvendando os mistérios do funcionamento da mente (GARCIA, 2013).

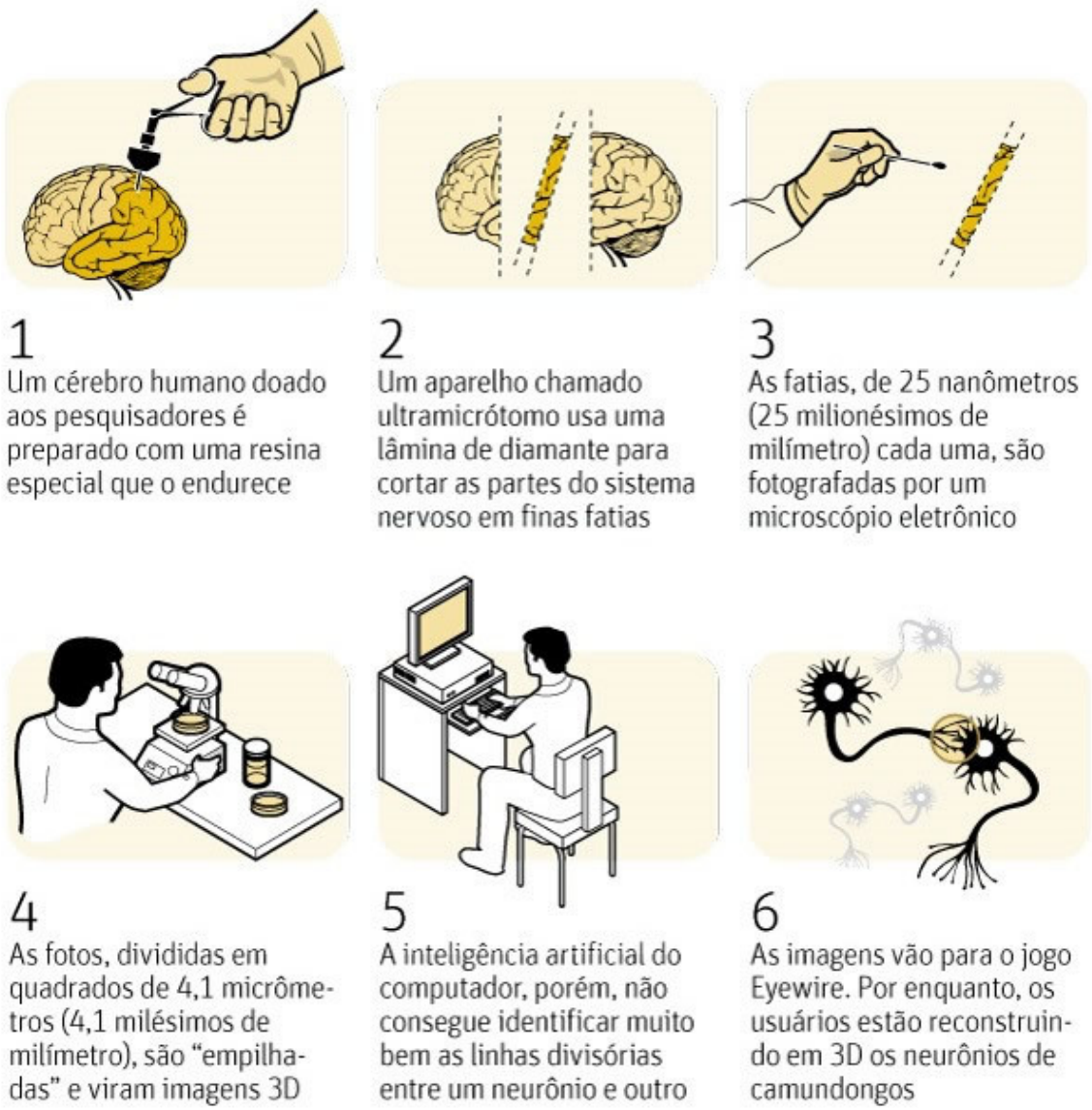
[2] https://eyewire.org/signup

Figura 16: Ambiente do jogo Eyewire (GARCIA, 2013)



Através de fotografias feitas por um microscópio (Figura 17), o jogo contém mais de um milhão de pacotes de imagens que através delas os jogadores ajudam a identificar os neurônios explorando um cubo e os pintam com ferramentas de desenho para delimitar seus formatos (GARCIA, 2013).

Figura 17: Como cientistas fotografam o cérebro (GARCIA, 2013)



Então, através de técnicas de inteligência artificial, os computadores descobrem como mapear os neurônios corretamente através dos jogadores e assim ajudando os cientistas a desvendarem os mistérios ainda não desvendados da mente (JAIN; SEUNG; TURAGA, 2010).

2.6.3 Calangos

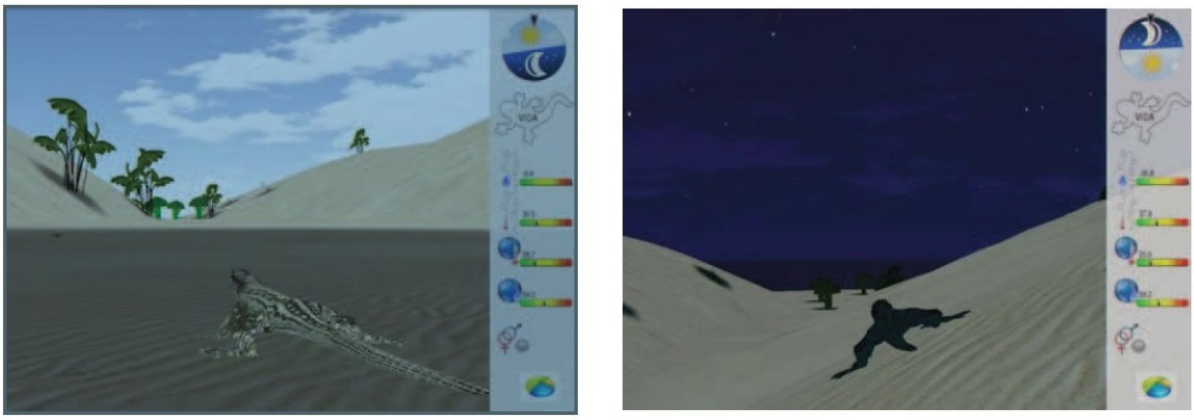
Calangos [3] é um jogo com enfoque na área da biologia para estimular a aprendizagem de alunos do ensino médio.

O jogo é baseado em uma modelagem ecológica no estado da Bahia, com o objetivo de possibilitar ao estudante um ambiente próximo do real e permitindo uma compreensão

[3] http://calangos.sourceforge.net/

de processos ecológicos e evolutivos (Figura 18), criando situações aos alunos de forma que sejam solicitados a relacionar fatores e dinâmicas que compõem o ecossistema ou até mesmo relacionar mecanismos de alteração no material genético, seleção natural e adaptação nas explicações sobre o surgimento de novas espécies de seres vivos (LOULA et al., 2011).

Figura 18: Capturas de telas do jogo Calangos em dois períodos simulados do dia, com a luz do sol e a noite. (LOULA et al., 2011)



O jogador passa-se por um lagarto e encontra-se em um ambiente contendo as características do habitat e do lagarto, onde será constantemente submetido à prova, e vários outros eventos com os quais terá de lidar, como os ataques de predadores ou a busca de alimento. De acordo com as decisões tomadas pelos jogadores, esses obstáculos poderão resultar em consequências tanto positivas quanto negativas ao metabolismo do réptil (LOULA et al., 2011).

Portanto, o Calangos é utilizado como ferramenta de apoio ao ensino e aprendizagem de ecologia e evolução no nível médio de escolaridade. Sendo assim, não se trata apenas de um jogo de entretenimento, também de aprendizagem decorrente da experiência na tentativa de resolver problemas reais.

2.6.4 Mineração de dados na classificação e seleção de Oncogenes

O projeto "Aplicação de Métodos Computacionais de Mineração de Dados na Classificação e Seleção de Oncogenes Medidos por Microarray", foi elaborado devido ao aumento de casos de câncer nos hospitais da rede pública, com isso seus criadores tiveram o objetivo de aplicar cinco métodos de mineração de dados na base de dados NVI60, construída com dados oriundos de experimentos de microarray contendo 1000 genes agrupados e nove classes de câncer (RODRIGUES; AMARAL, 2012).

Os métodos de mineração utilizados no trabalho foram: J48, Random Forest, PART,

IBK e Naive Bayes, pertencentes ao ambiente Weka [4] onde constataram ao longo de suas pesquisas que o IBK obteve a melhor classificação, enquanto o J48 e PART conseguiram diminuir o conjunto de genes drasticamente.

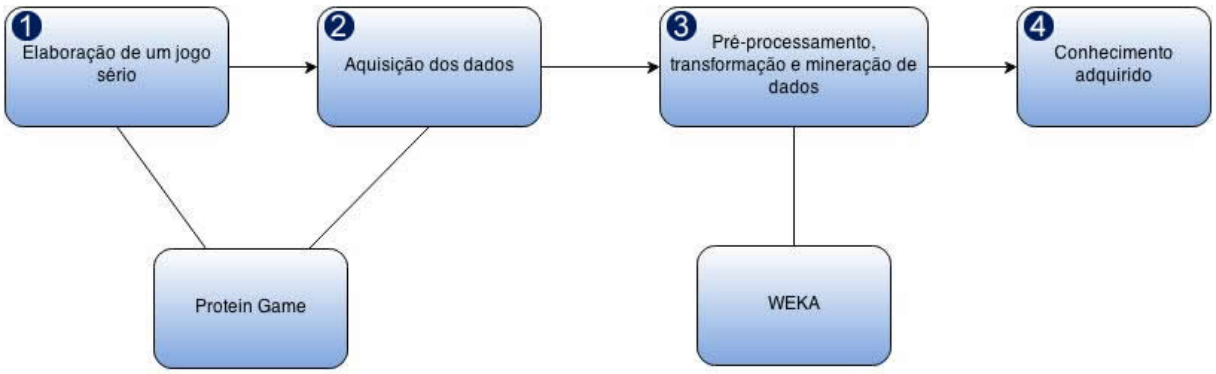
Os resultados foram obtido através de cálculos utilizando equações que retornam o *score* obtido por cada método de mineração de dados. Com isso, através dos resultados adquiridos, os mesmos podem ser utilizados como ferramentas que visam a auxiliar no enfrentamento do câncer, podendo classificar novos casos as relações gene/gene e gene/câncer sendo assim facilitando cientistas a combater o câncer (RODRIGUES; AMARAL, 2012).

[4] http://www.cs.waikato.ac.nz/ml/weka/book.html

3 Materiais e Métodos

Este capítulo apresenta os materiais (software, equipamento, especialista) que foram utilizados e como cada etapa apresentadas na Figura 19, foi desenvolvida.

Figura 19: Fluxograma com as ferramentas utilizadas



3.1 Implementação de um jogo sério biológico

Para coletar informações da inteligência humana, uma alternativa existente é de elaborar um jogo para a solução de problemas biológicos. Com ênfase no objetivo da pesquisa, foi elaborado como ferramenta, um jogo voltado para a área da biologia, mais precisamente ao dobramento de proteínas no modelo HP. O objetivo do jogo consiste em despertar o interesse dos jogadores para que se possa obter um grande número de dobramentos fornecidos pelos mesmos. Para isso, é necessária a utilização de linguagens de programação que possibilitem efetuar jogos através da Internet, fazendo com que se tenha um acesso facilitado.

3.1.1 Implementação

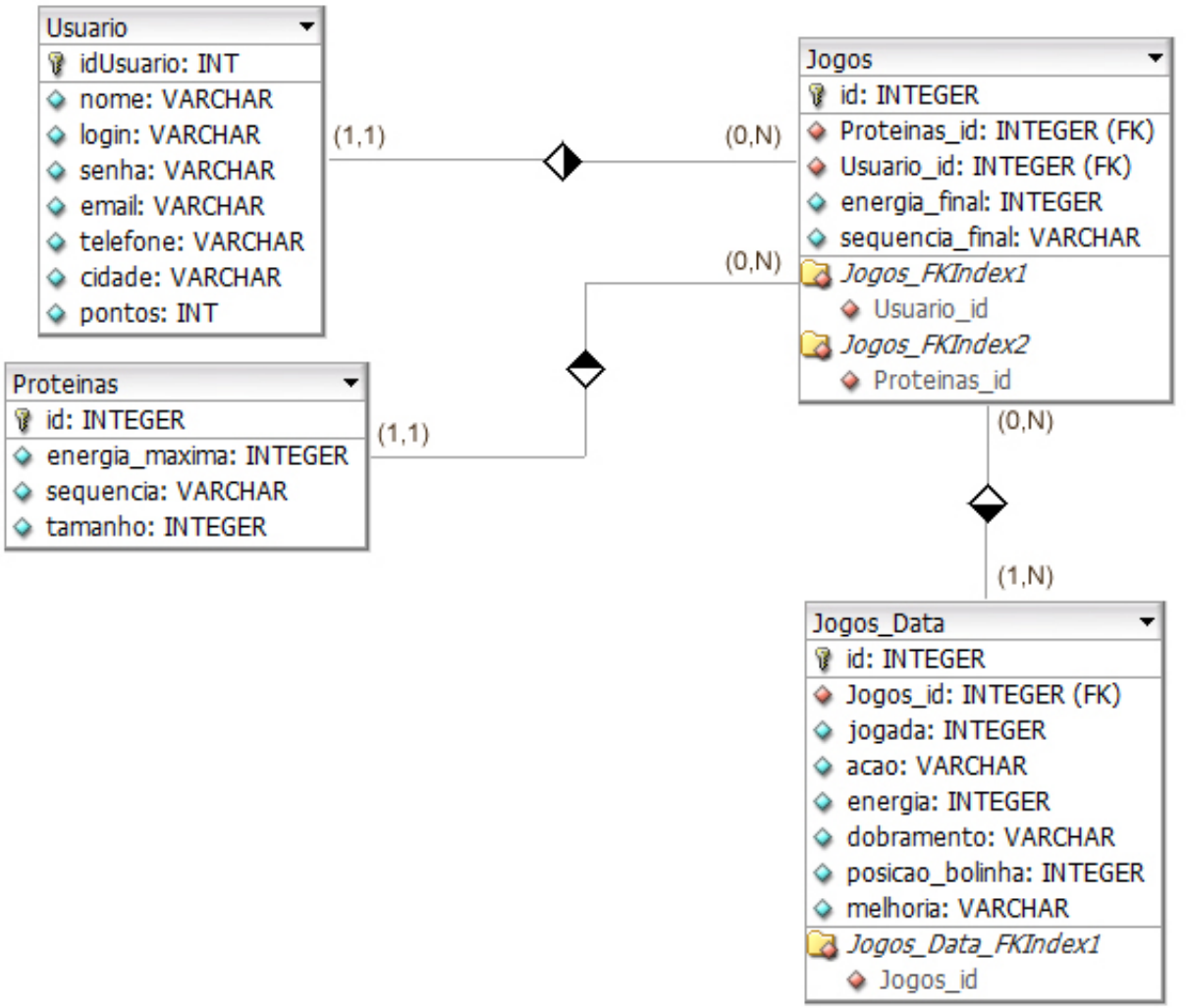
Foram utilizadas as linguagens HTML5 juntamente com o CSS, para construção de páginas, possibilitando confeccionar a interface do jogo, fazendo com que funcione em qualquer navegador e que possa ser hospedado em qualquer servidor.

Também foi utilizado as linguagens javascript, que faz o papel de *frontend* interagindo na interface dos usuários, agindo juntamente com técnicas Ajax, para interagir com o *backend* sem influenciar na jogabilidade e na interface do jogo e o PHP(*Personal Home Page*) utilizado para o desenvolvimento de aplicações presentes e atuantes no lado do servidor agindo como *backend*, em conjunto com o MYSQL que é um sistema de gerenciamento de banco de dados (SGBD) salvando todos os dados necessários no banco de dados hospedado juntamente com o jogo.

3.1.2 Banco de dados

Na Figura 20 pode-se ver o modelo relacional da estrutura do banco de dados que contém quatro tabelas com suas relações e atributos, seguido de suas especificações.

Figura 20: Modelo de relacionamento do banco de dados



- **Tabela Usuario:** utilizada para o armazenado dos dados de cadastro dos jogadores no banco. Ela também possui um relacionamento de 1:N com a Tabela Jogos, onde um jogador contém vários jogos e um jogo contém apenas um jogador.
- **Tabela Proteinas:** armazena as proteínas que são fornecidas aos jogadores. Cada estrutura possui seu tamanho, sua sequência HP e sua energia mínima. Esta Tabela contém um relacionamento de 1:N com a Tabela Jogos, onde uma proteína pode ter zero ou N jogos e um jogo contém uma proteína.
- **Tabela Jogos:** A cada jogo feito pelos usuários, é criada uma nova entrada nessa Tabela, contendo os dados necessários para relacionar o jogo com o usuário e também

todas as jogadas feitas. Esta Tabela contém um relacionamento de 1:N com a Tabela Jogos_Data, onde um jogo pode ter zero ou N jogadas e uma jogada contém apenas um jogo.

- **Tabela Jogos_Data:** Esta Tabela é a grande ferramenta para a pesquisa, pois nela é armazenada todos os tipos de jogadas que os usuários fizeram em todos seus jogos. Com isso é possível saber qual o número da jogada, a ação feita nesta jogada, a energia atingida na jogada, o dobramento em que se encontra a estrutura, a posição do aminoácido em que foi feito o dobramento e também uma comparação entre a jogada atual e a anterior, classificando-a como uma jogada que fez uma energia melhor, pior ou igual.

  Os dados armazenados nesta Tabela são utilizados nas técnicas de mineração de dados.

3.1.3 Interface

Na Figura 21, pode-se analisar a interface do website do jogo, que contém os atalhos para diferentes páginas. O botão Ranking contém uma Tabela com o ranking dos jogadores segundo suas pontuações. No botão de "Regras", está disponível um conteúdo onde os visitantes possam aprender as funcionalidades do jogo. No botão "Sobre nós", foi elaborada para apresentar a equipe e o objetivo científico do jogo. E o botão de "Contato", Existe um formulário onde os visitantes e jogadores possam entrar em contato com a equipe como sugestões, criticas e duvidas.

Figura 21: Interface do jogo

Como é mostrado também na Figura 21, é possível efetuar um cadastro que fornece o acesso ao jogo. Sendo assim, ao efetuar o login, os usuários tem acesso a uma página como mostra a Figura 22 que fornece as opções Novo Jogo, Continuar jogo Salvo, Perfil, Alterar Cadastro e Sair.

Ao clicar no botão Novo Jogo, os jogadores podem escolher a proteína que desejam jogar através de uma lista que contém todas as estruturas das proteínas fornecidas no banco de dados conforme a Figura 23.

Figura 22: Interface - Usuário logado

Figura 23: Lista das proteínas que podem ser selecionadas para jogar

Após selecionar a proteína que deseja, o jogo contém uma interface como mostra na Figura 24 onde:

1. Mostra a quantidade de energia em que a estrutura se encontra
2. Botão que possibilita o jogador a finalizar o jogo
3. Botão onde possibilita o jogador salvar o jogo caso queira continuar a partida que deseja
4. Botão de ajuda onde mostra uma documentação de como efetuar um dobramento e etc.
5. A estrutura da proteína modelo HP esticada, que é sua posição inicial dentro do jogo

   Do item 6 ao item 10, são ferramentas que foram criadas para facilitar o jogador possibilitando efetuar uma sequência de dobramento mais rapidamente:

6. É uma ferramenta que possibilita o jogador selecionar um intervalo de aminoácidos da estrutura e efetuar um dobramento igual ao da Figura 25a
7. Possibilita o jogador a espelhar a estrutura da proteína, como mostra a Figura 25b

8. Possibilita o jogador a selecionar um intervalo de aminoácidos fazendo com que todos os aminoácidos deste intervalo fiquem alinhados, como mostra a Figura 25c

9. Caso o jogador não goste da ultima jogada feita, essa ferramenta possibilita voltar a estrutura ao seu estado anterior

10. Esta ferramenta faz com que a estrutura da proteína volte ao seu estado inicial como mostra a Figura 25d.

Figura 24: Interface - jogo iniciado

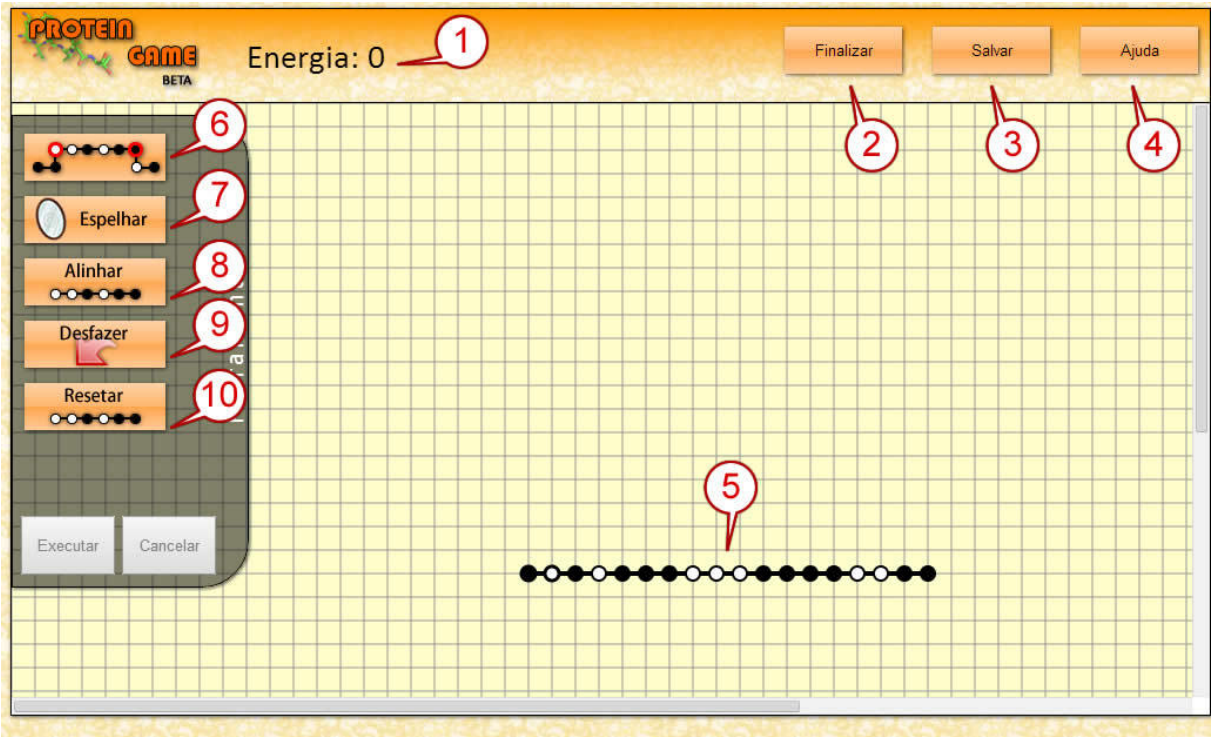
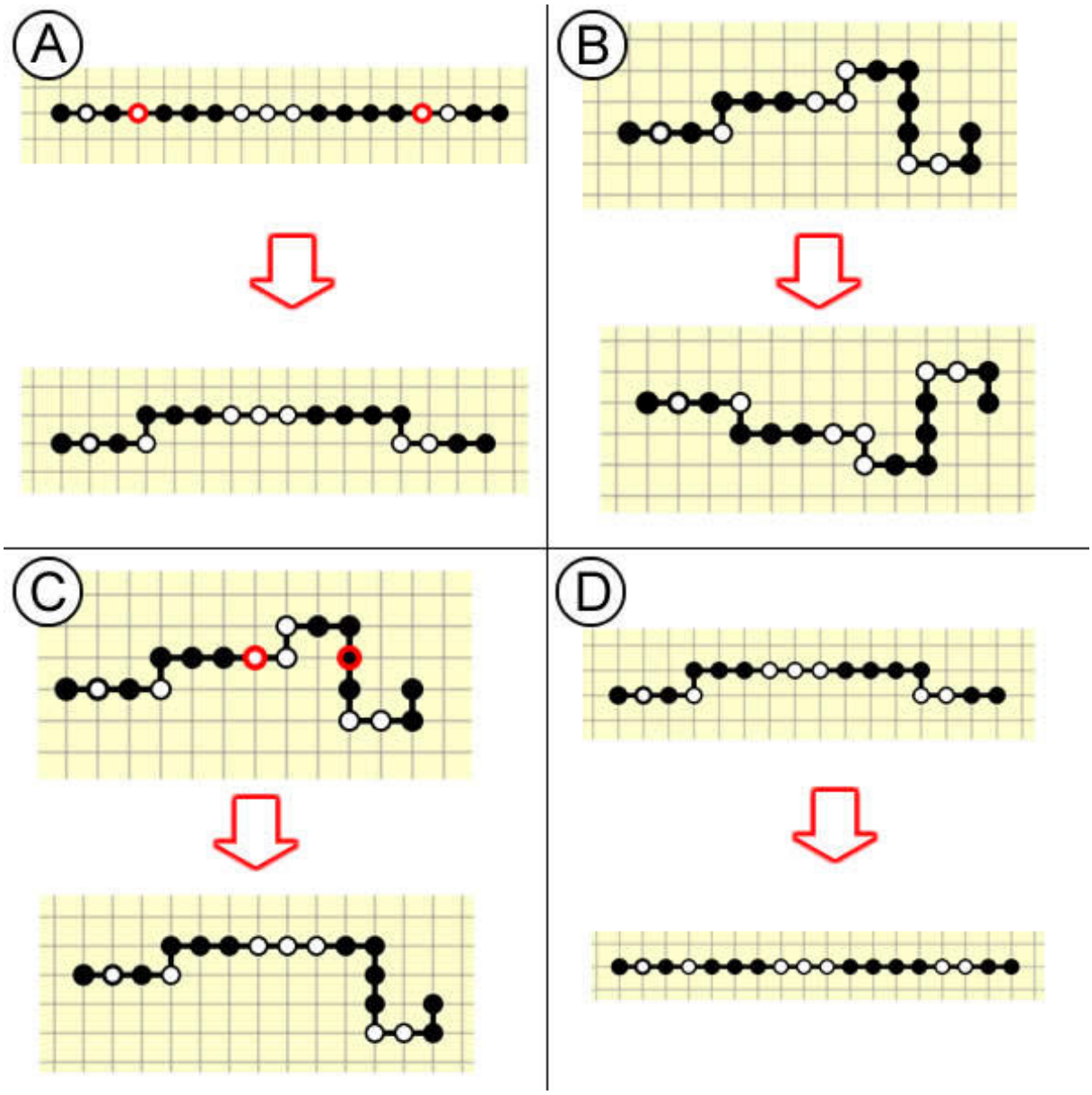


Figura 25: Exemplos do uso das ferramentas

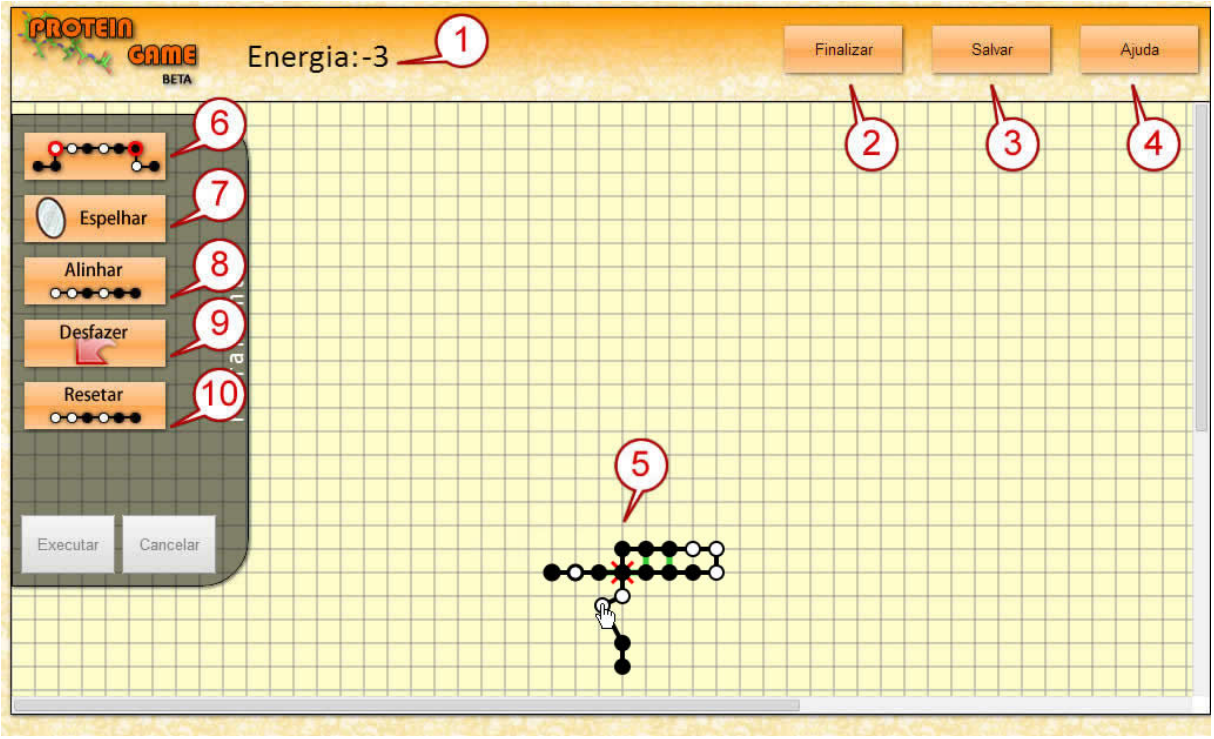


O jogo foi desenvolvido para que se tenha jogabilidade simples e intuitiva: Portanto, cada aminoácido da estrutura da proteína modelo HP que é simbolizada por círculos brancos e pretos, são arrastáveis através da função do mouse clicando e segurando. O círculo clicado se movimenta juntamente com o cursor do mouse e logo após o jogador soltá-lo, a estrutura irá se dobrar automaticamente ajustando os outros círculos nos nós da grade e com as suas ligações do tamanho de um quadrante da grade.

A cada ponto de energia que gera no dobramento da estrutura, a interface simboliza os contatos hidrofóbicos, com um risco verde, como mostra no item 5 da Figura 26 e também atualiza a quantidade de energia no placar, mostrado no item 1 da Figura 26. Caso o jogador efetue dobramentos que cruze a estrutura da proteína, será colocado um "x"vermelho indicando

este cruzamento e a energia será zerada no placar.

Figura 26: Interface - Ambiente do jogo com a estrutura dobrada



3.2 Aquisição dos Dados

Após feito os últimos ajustes na implementação do jogo, este foi testado com 40 alunos de graduação no horário de aula de professores que cederam uma parte do seu horário. Para que seja possível o teste. Primeiramente foi feita uma breve introdução a respeito do projeto e algumas explicações de como funciona o jogo. Também foi priorizado nas aplicações do jogo, as proteínas S1 (HPHPHHHPPPHHHHPPHH) e S2 (PHPPHPHHHPHHPHHHHH) para que fosse possível obter um volume maior dessas duas estruturas, para os testes.

Ao todo, foi possível aplicar o jogo em duas aulas, com uma duração de uma hora. Os alunos também foram incentivados a jogar em outros horários e que recomendassem para seus colegas.

No total, 44 jogadores foram registrados em nosso banco de dados, obtendo um valor de 401 jogos criados, totalizando 12059 jogadas.

3.3 Pré-processamento, transformação e mineração dos dados

Para que seja possível chegar a resultados relevantes através de técnicas de mineração de dados, é importante preparar os dados obtidos e conforme são submetidos a testes, gera-

se conhecimento com os quais é possível identificar novas formas de fornecer as entradas e submeter a novos testes até chegar-se a resultados satisfatórios considerando as taxas de acerto da classificação e as regras geradas. A seguir será explicado a respeito de todos os testes executados e suas respectivas características de dados que foram submetidos, bem como qual a técnica de mineração utilizada.

A quantidade de instâncias por folha para cada um dos testes foi definida proporcionalmente a quantidade de instâncias utilizadas em cada teste. Por exemplo, para o teste 1 onde tem-se uma quantidade maior de instâncias utilizou-se um mínimo de 41 instâncias por folha onde é possível uma restrição maior do que no teste 3 onde tem-se uma quantidade menor de instâncias e portanto foi executado com 19 instâncias por folha, como apresentada a Tabela 3.

3.3.1 Weka e Algoritmo J48

Os testes foram realizados com ajuda do software WEKA que através de um arquivo CSV (Figura 27) possibilita utilizar os dados exportados do banco de dados do jogo.

Figura 27: Exemplo de um arquivo CSV

```
idproteina,idjogo,jogada,dobramento,energia,movimento,posicao,melhoria
1,97,0,FFFFFFFFFFFFFFFF,0,,0,diminuiu
1,97,1,FFFFFFFFEFFFFFFF,0,E,9,igual
1,97,2,FFFFFFFFEFEFFFFF,0,E,11,igual
1,97,3,FFFFFFFFEFFFFFFF,0,U,0,igual
1,97,4,FEDFFFFDEEFFFFFF,0,Q,0,igual
1,97,5,FEDFFFFEDDFFFFFF,0,E,0,igual
1,97,6,FEDFFFFFFFFFFFFF,0,L,0,igual
1,97,7,FFFFFFFFFFFFFFFF,0,R,0,igual
1,97,8,FFFFFEFFFFFFFFFF,0,E,5,igual
1,97,9,FFFFFEEFFFFFFFFF,0,E,6,igual
1,97,10,FFDFFEEFFFFFFFFF,0,D,2,igual
1,97,12,FFEFFDEDFFDEFFFF,0,Q,0,diminuiu
1,97,13,FFEFFEDEFFEEFFFF,2,E,0,diminuiu
1,97,14,FFEFFDEDFFDEFFFF,0,U,0,melhorou
1,97,15,FFEFFDEDFFDEFFFF,0,Q,0,igual
1,97,16,FFEFFDEDFFDEFFFF,0,Q,0,igual
1,97,17,FFEFFDEDFFDDFFFF,2,E,0,diminuiu
1,97,18,FFEFFDEDFFDEFFFF,0,U,0,melhorou
1,97,19,FFEFFDEDFFDEDFFE,0,Q,0,igual
1,97,20,FFEFFDEDFFDEFFFF,0,U,0,igual
1,97,21,FFEFFDEDFFDEEDFE,0,Q,0,igual
1,97,22,FFEFFDEDFFDEEDDE,0,D,14,igual
1,97,23,FFEFFDEDFFDEEDDE,0,E,15,igual
1,97,24,FFFFFFFFFFFFFFFF,0,R,0,igual
```

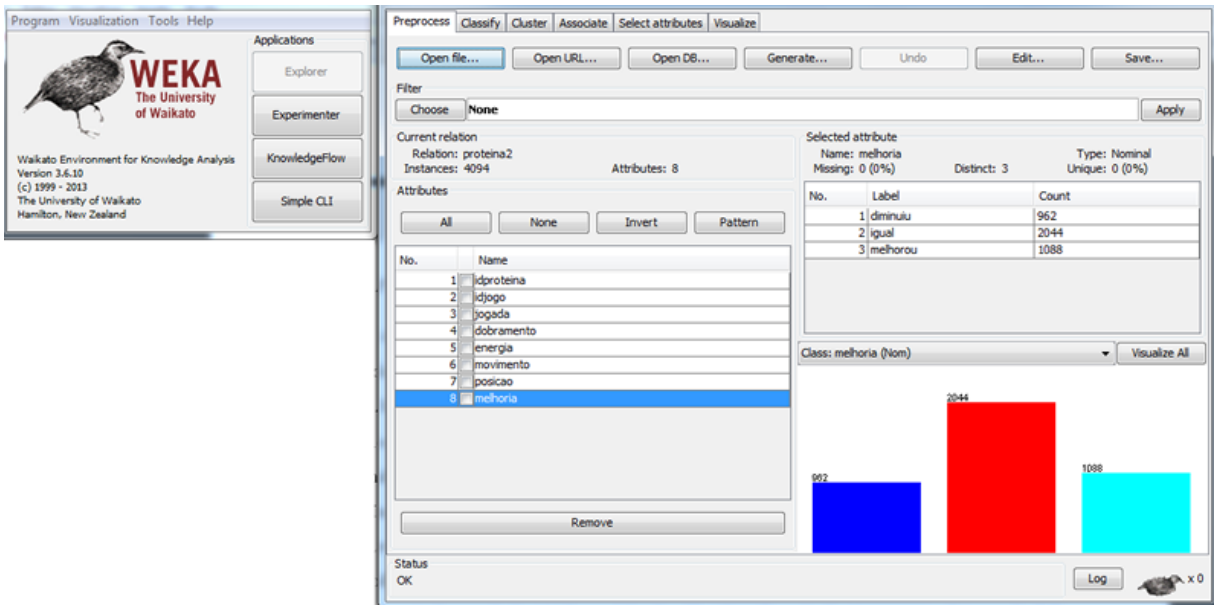
O WEKA [1] é um software gratuito desenvolvido pela Universidade de Waikato na Nova Zelândia. Surgiu da necessidade de um ambiente de trabalho que possibilitem aos pesquisadores obter um fácil acesso as técnicas mais recentes em aprendizado de máquina. No surgimento

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

do projeto em 1992, os algoritmos de aprendizado estavam disponíveis em diversas linguagens de programação para o uso de diversas plataformas e o operava em uma variedade de formatos de dados. A tarefa de coletar os esquemas de aprendizado para estudos comparativos em uma coleção de dados é relativamente complexa. Com esta visão, o WEKA não só forneceria o conjunto de ferramentas de algoritmos de aprendizado, mas também um ambiente onde os pesquisadores poderiam implementar novos algoritmos sem se preocupar com a infraestrutura para manipulação dos dados e validação dos modelos (HALL et al., 2009).

Weka tem uma interface simples (Figura 28) que facilita o trabalho do pesquisador, onde facilmente pode-se fornecer as entradas com os dados através de arquivos de diferentes formatos. Após a leitura das entradas é possível selecionar as colunas que o pesquisador deseja trabalhar e através de abas é possível selecionar o tipo de técnica de mineração de dados e também diversos tipos de algoritmos dessas técnicas. O algoritmo de mineração de dados utilizado foi o J48 que foi utilizado como exemplo e explicado detalhadamente na seção 2.5.1.

Figura 28: Interface do software WEKA



O arquivo de saída CSV adquirido ao exportar os dados do banco de dados do jogo é semelhante ao da Figura 27. Este arquivo é um arquivo texto comum, que funciona como uma Tabela com linhas e colunas, mas as colunas são separadas por vírgula e as linhas separadas por um "Enter"(nova linha). Na primeira linha, contém os atributos separados por vírgula e que serão interpretados pelo WEKA e cada linha restante é um dobramento feito na estrutura por intermédio do jogo.

Cada coluna do arquivo é interpretada pelo WEKA como um atributo, portanto será explicado abaixo as características de cada coluna(atributo):

1. **Jogada:** Contém o número da jogada de todos os jogos
2. **Dobramento:** Contém estado da estrutura da proteína possibilitando saber como a estrutura está dobrada
3. **Energia:** Contém a quantidade de energia que a estrutura da proteína está gerando no estado que se encontra
4. **Movimento:** Contém o movimento feito na jogada, podendo ser: E(Esquerda), D(Direita), U(Ferramenta Desfazer), Q(Ferramenta que efetua o dobramento mostrado na Figura 25a), Z(Ferramenta que efetua o espelhamento da estrutura como mostrado na Figura 25b), L(Ferramenta que efetua o alinhamento do intervalo de aminoácidos selecionados como mostrado na Figura 25c) e a L(Ferramenta que reseta a estrutura deixando em sua posição inicial como mostrado na Figura 25d)
5. **Posicao:** Contém a posição do aminoácido que foi efetuado o dobramento
6. **Melhoria:** Mostra se a energia gerada com o estado em que a estrutura se encontra melhorou, piorou ou se manteve igual referente ao estado anterior

Os dados adquiridos dessas proteínas foram submetidos ao algoritmo J48 do WEKA para descobrir possíveis técnicas de dobramento dessas proteínas. A seguir será explicado detalhadamente como foram feitos os testes.

3.3.2 Características dos testes

A Tabela 3 apresenta o número de instâncias e o número mínimo de instâncias por folha, para os testes realizados.

Tabela 3: Nº de Instâncias e Nº mínimo de instâncias por folha

Teste	Nº de Instâncias utilizadas	Nº mínimo de instâncias por folha
Teste 1	4.148	41
Testes 2	3.910	39
Testes 3	1.940	19
Testes 4	2.960	29
Testes 5	2.268	22

O número total de instâncias são todas as jogadas adquiridas. Nos testes realizados, tem-se números de instâncias utilizadas de quantidade diferente, pois foram realizados pré-processamentos na base de dados adquirida.

Os números mínimos de instâncias por folha utilizados seguem proporcionalmente o número de intâncias que seriam submetidas ao teste.

Portanto foi mantido os valores mínimos de instâncias por folha de 1% do número total de intâncias. No teste 1 foi utilizado 4.148 instâncias e um mínimo de 41 instâncias por folha. Nos testes 2 foram utilizados 3.910 instâncias e um mínimo de 39 instâncias por folha. Nos testes 3 foram utilizadas 1.940 instâncias e um mínimo de 19 instâncias por folha. E nos testes 4 foram utilizadas 2.960 instâncias e um mínimo de 29 instâncias por folha.

3.3.3 Teste 1

Para este teste foi gerado o arquivo CSV contendo todas as jogadas de todos os jogos executados apenas da proteína S1 (HPHPHHHPPPHHHHPPHH), utilizando um total de 4.148 instâncias, para realizar análise de como a classificação irá se comportar mediante as instâncias fornecidas. E foram classificadas contendo os atributos: jogada, movimento, posição e melhoria com um número mínimo de 41 instâncias por folha e seu atributo classificador foi o melhoria.

3.3.4 Testes 2

Para este teste foi gerado o arquivo CSV contendo as jogadas de todos os jogos executados apenas da proteína S1 (HPHPHHHPPPHHHHPPHH). A jogada zero foi excluída porque é a jogada inicial em todos os jogos e a jogada 1 foi concluído porque a proteína fica em um ângulo de 90 graus, não alterando sua energia portanto também foi ignorada. Este teste conteve um total de 3.910 instâncias e teve como seu atributo classificador o atributo melhoria e um número mínimo de 39 instâncias por folha.

As instâncias foram classificadas de duas formas diferentes contendo diferentes atributos:

1. Os atributos utilizados neste teste foram: jogada, movimento, posição e melhoria.
2. Os atributos utilizados neste teste foram: movimento, posição e melhoria.

3.3.5 Testes 3

Para este teste foi gerado o arquivo CSV contendo as jogadas de todos os jogos executados apenas da proteína S1 (HPHPHHHPPPHHHHPPHH). A jogada zero foi excluída porque é a jogada inicial em todos os jogos e a jogada 1 foi concluído que a proteína fica em um ângulo de 90 graus, não alterando sua energia portanto também foi ignorada, e também foram ignoradas todas as jogadas de melhoria = *igual*. Dessa forma, buscou-se balancear a quantidade de instâncias da base de conhecimento utilizado. Este teste conteve um total

de 1.940 instâncias e teve como seu atributo classificador o atributo melhoria e um número mínimo de 19 instâncias por folha.

As instâncias foram classificadas de duas formas diferentes contendo diferentes atributos:

1. Os atributos utilizados neste teste foram: jogada, movimento, posição e melhoria.
2. Os atributos utilizados neste teste foram: movimento, posição e melhoria.

3.3.6 Testes 4

Para este teste foi gerado o arquivo CSV contendo as jogadas de todos os jogos executados apenas da proteína S1 (HPHPHHHPPPHHHHPPHH). A jogada zero foi excluída porque é a jogada inicial em todos os jogos e a jogada 1 foi concluído porque a proteína fica em um ângulo de 90 graus, não alterando sua energia, portanto também foi ignorada, e também foram ignoradas aleatoriamente 920 jogadas de melhoria = *igual*. Dessa forma buscou-se balancear a quantidade de instâncias da base de conhecimento utilizado. Este teste conteve um total de 2.960 instâncias e teve como seu atributo classificador o atributo melhoria e um número mínimo de 29 instâncias por folha.

As instâncias foram classificadas de duas formas diferentes contendo diferentes atributos:

1. Os atributos utilizados neste teste foram: jogada, movimento, posição e melhoria.
2. Os atributos utilizados neste teste foram: movimento, posição e melhoria.

3.3.7 Testes 5

Para este teste foi gerado o arquivo CSV contendo as jogadas de todos os jogos executados apenas da proteína S1 (HPHPHHHPPPHHHHPPHH). A jogada zero foi excluída porque é a jogada inicial em todos os jogos e a jogada 1 foi concluído porque a proteína fica em um ângulo de 90 graus, não alterando sua energia, portanto também foi ignorada, e também foram ignoradas 920 jogadas onde melhoria = *igual* e todas as jogadas de melhoria = *piorou*, pois foi identificado que as classificações estão tendo um alto índice de erros ao classificar as melhorias iguais a piorou. Dessa forma, buscou-se balancear a quantidade de instâncias da base de conhecimento utilizado. Este teste conteve um total de 2.268 instâncias e teve como seu atributo classificador o atributo melhoria e um número mínimo de 22 instâncias por folha.

As instâncias foram classificadas de duas formas diferentes contendo diferentes atributos:

1. Os atributos utilizados neste teste foram: jogada, movimento, posição e melhoria.

2. Os atributos utilizados neste teste foram: movimento, posição e melhoria.

4 Resultados

A definição de todos os parâmetros utilizados para os testes foram apresentados na seção 3.3. Para facilitar o entendimento os atributos utilizados, estes serão novamente citados neste capítulo.

4.1 Percentuais de instâncias corretamente classificadas

Foram realizados 5 testes no conjunto de dados obtidos. A Tabela 4 apresenta percentuais de intâncias classificadas corretamente.

Tabela 4: Percentuais de instâncias corretamente classificadas dos 5 testes realizados

Teste	Acurácia para Cross Validation
Teste 1	60.29%
Teste 2 versão 1	56.24%
Teste 2 versão 2	54.52%
Teste 3 versão 1	67.47%
Teste 3 versão 2	68.04%
Teste 4 versão 1	52.97%
Teste 4 versão 2	53.10%
Teste 5 versão 1	67.98%
Teste 5 versão 2	67.63%

Pode-se verificar que os percentuais de acerto, para todos os testes, foram superiores a 50%. O teste 3 versão 2 obteve o maior percentual, 68,04%. Nas seções subsequentes são apresentados os testes e explicados os resultados obtidos.

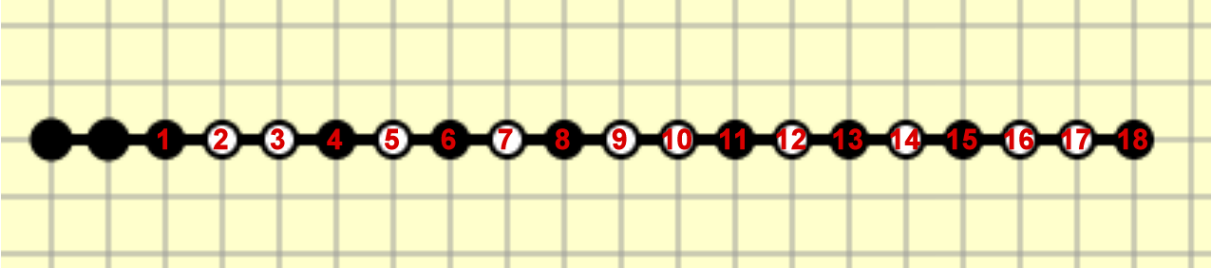
Além de analisar o percentual de acertos e matriz de confusão, será analisado também possíveis regras de dobramento de proteína e em alguns momentos será abordado a respeito das melhorias e as posições dos aminoácidos.

Conforme o modelo HP, o objetivo maior de um dobramento da estrutura é atingir um menor número de energia possível. Portanto, é importante frisar que ao observar que a energia *melhorou* significa que a estrutura em sua posição atual obteve um menor número de energia, sendo assim, obtendo uma melhora energética e ao dizer que a energia *piorou*, significa que a energia aumentou, piorando seu valor energético.

Em alguns momentos das análises dos resultados, será verificado as posições dos aminoácidos em que os jogadores executaram o dobramento. Na Figura 29 pode-se ver um exem-

plo das posições, onde a posição 1 começa no terceiro aminoácido devido a fixação dos dois primeiros aminoácidos que não fariam diferença nos resultados.

Figura 29: Exemplo das posições dos aminoácidos na estrutura da proteína modelo HP



4.2 Teste 1

As instâncias fornecidas neste teste foram classificadas contendo os atributos: jogada, movimento, posição e melhoria, com um número mínimo de 41 instâncias por folha e seu atributo classificador foi melhoria.

Neste teste foi obtida uma árvore classificatória como mostra a Figura 30 que contém uma taxa de acerto de 60.29% e uma matriz de confusão como mostra na Tabela 5[1].

Tabela 5: Matriz de confusão teste 1

	A	B	C
A = piorou	**276**	527	166
B = igual	92	**1701**	288
C = melhorou	66	562	**470**

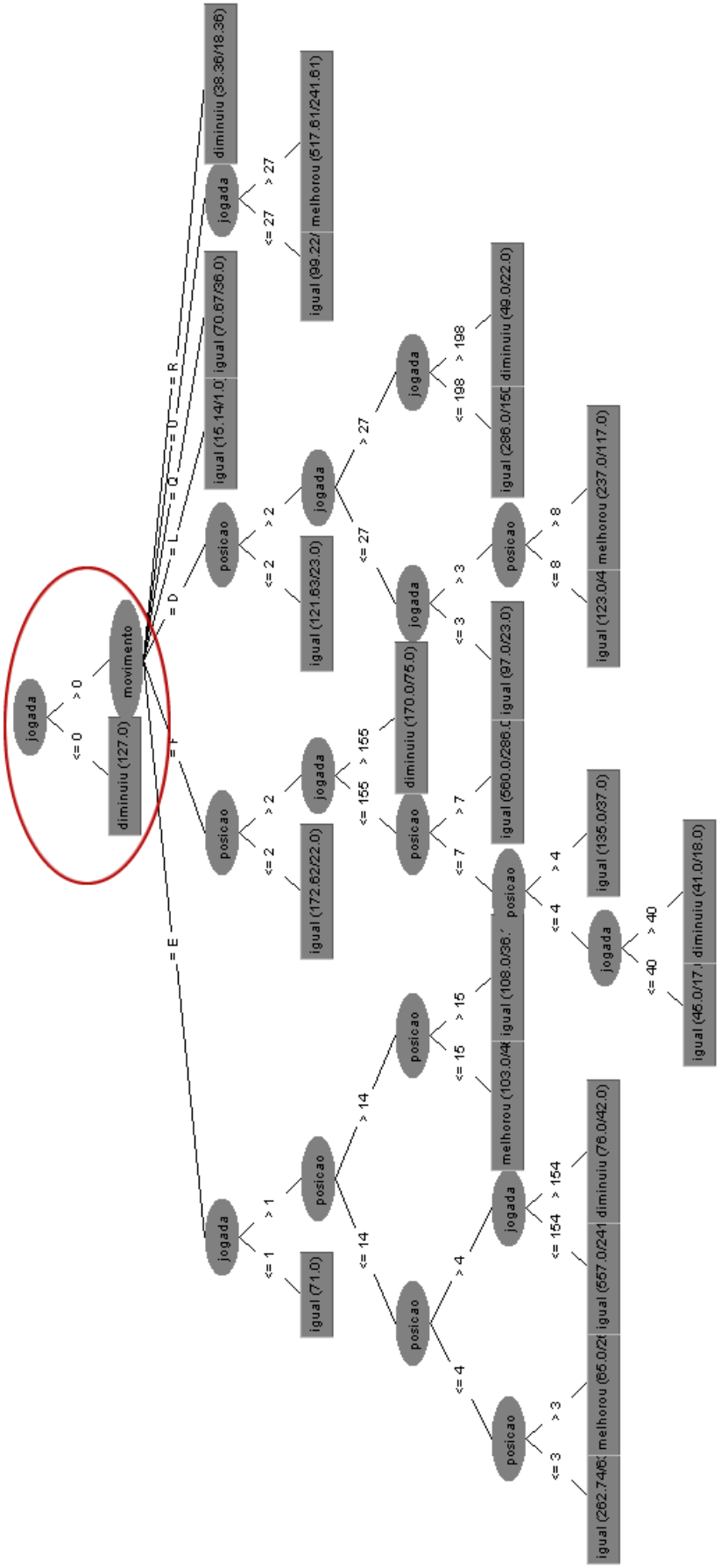
Conforme a Tabela 5, pode-se observar que:

- As instâncias com melhoria = *piorou* obteve 276 instâncias classificadas corretamente
- As instâncias com melhoria = *igual* obteve 1701 instâncias classificadas corretamente
- As instâncias com melhoria = *melhorou* obteve 470 instâncias classificadas corretamente

Pelos dados obtidos, pode-se concluir que a primeira ramificação, ou seja, a jogada zero que contém a estrutura inicial, bem como a jogada um que contém apenas um dobramento, não influenciam na melhoria ou no valor da energia. Portanto, 127 instâncias não estão trazendo nenhum conhecimento para esta classificação.

[1] Em todas as matrizes de confusão dos testes realizados, a matriz principal foi destacada em negrito, para facilitar a leitura das instâncias corretamente classificadas.

Figura 30: Árvore de classificação gerada no teste 1



Desta forma, nos próximos testes as jogadas zero e um não serão utilizadas. Assim, este teste acabou se tornando um pré-processamento de dados, pois reduziu a base de conhecimento para os testes subsequentes.

4.3 Testes 2

- **Versão 1**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: jogada, movimento, posição e melhoria com um número mínimo de 39 instâncias por folha e seu atributo classificador foi o melhoria.

  Neste teste, foi obtido uma árvore classificatória como mostra a Figura 31 que contém uma taxa de 56.24% de acerto e uma matriz de confusão como mostra na Tabela 6.

  Tabela 6: Matriz de confusão teste 2 versão 1

	A	B	C
A = piorou	**180**	508	154
B = igual	127	**1542**	301
C = melhorou	544	77	**477**

  Conforme a Tabela 6, pode-se observar que:

  - As instâncias com melhoria = *igual* obteve 1542 instâncias classificadas corretamente
  - As instâncias com melhoria = *melhorou* obteve 477 instâncias classificadas corretamente
  - As instâncias com melhoria = *piorou* obteve 180 instâncias classificadas corretamente

  Pode-se observar que na Figura 31 existem caminhos classificatórios que foram marcados com diferentes cores devido a relevância de seus números de instâncias classificadas:

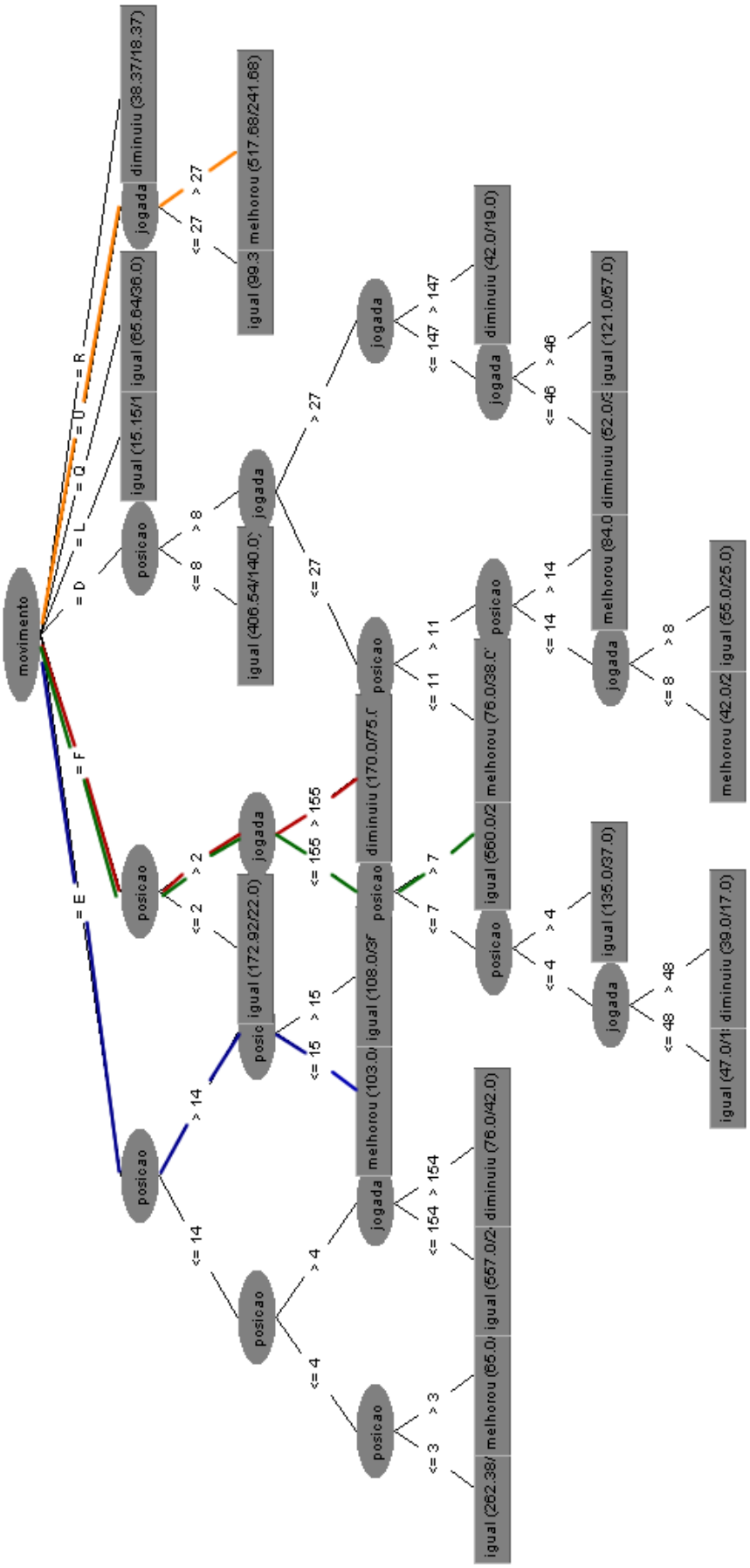
  Caminho Vermelho: Este caminho mostra que com o movimento F(frente) nos aminoácidos maiores que dois nas jogadas > 155, a energia da estrutura diminui.

  Caminho Azul: Este caminho mostra que efetuando um movimento E(esquerda) nos aminoácidos 14 ou 15 a energia da estrutura melhora.

  Caminho Verde: Este caminho mostra que deixando F(frente) nos aminoácidos > 7 nas jogadas $<= 155$, a energia da estrutura permanece igual.

  Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar) nas jogadas > 27.

Figura 31: Árvore de classificação gerada no teste 2 versão 1



Apesar de achar possíveis regras, este teste não teve uma taxa de acerto muito boa e pode-se observar que na matriz de confusão (Tabela 6) obteve uma grande desigualdade da classificação "igual"para as outras. Portanto, os próximos testes irão buscar melhorar o percentual de acerto e a matriz de confusão.

- **Versão 2**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: movimento, posição e melhoria com um número mínimo de 39 instâncias por folha e seu atributo classificador foi o "melhoria". Foram descartadas as jogadas de número zero e número um.

  Neste teste, foi obtido uma árvore classificatória como mostra a Figura 32 que contém uma taxa de 54.52% de acerto e uma matriz de confusão como mostra na Tabela 7.

Tabela 7: Matriz de confusão teste 2 versão 2

	A	B	C
A = piorou	**79**	220	543
B = igual	88	**1509**	373
C = melhorou	25	529	**544**

Conforme a Tabela 7, pode-se observar que:

- As instâncias com melhoria = *igual* obteve 1509 instâncias classificadas corretamente
- As instâncias com melhoria = *melhorou* obteve 544 instâncias classificadas corretamente
- As instâncias com melhoria = *piorou* obteve 79 instâncias classificadas corretamente

Pode-se observar que na Figura 32 existem caminhos classificatórios que foram marcados com diferentes cores devido a relevância de seus números de instâncias classificadas:

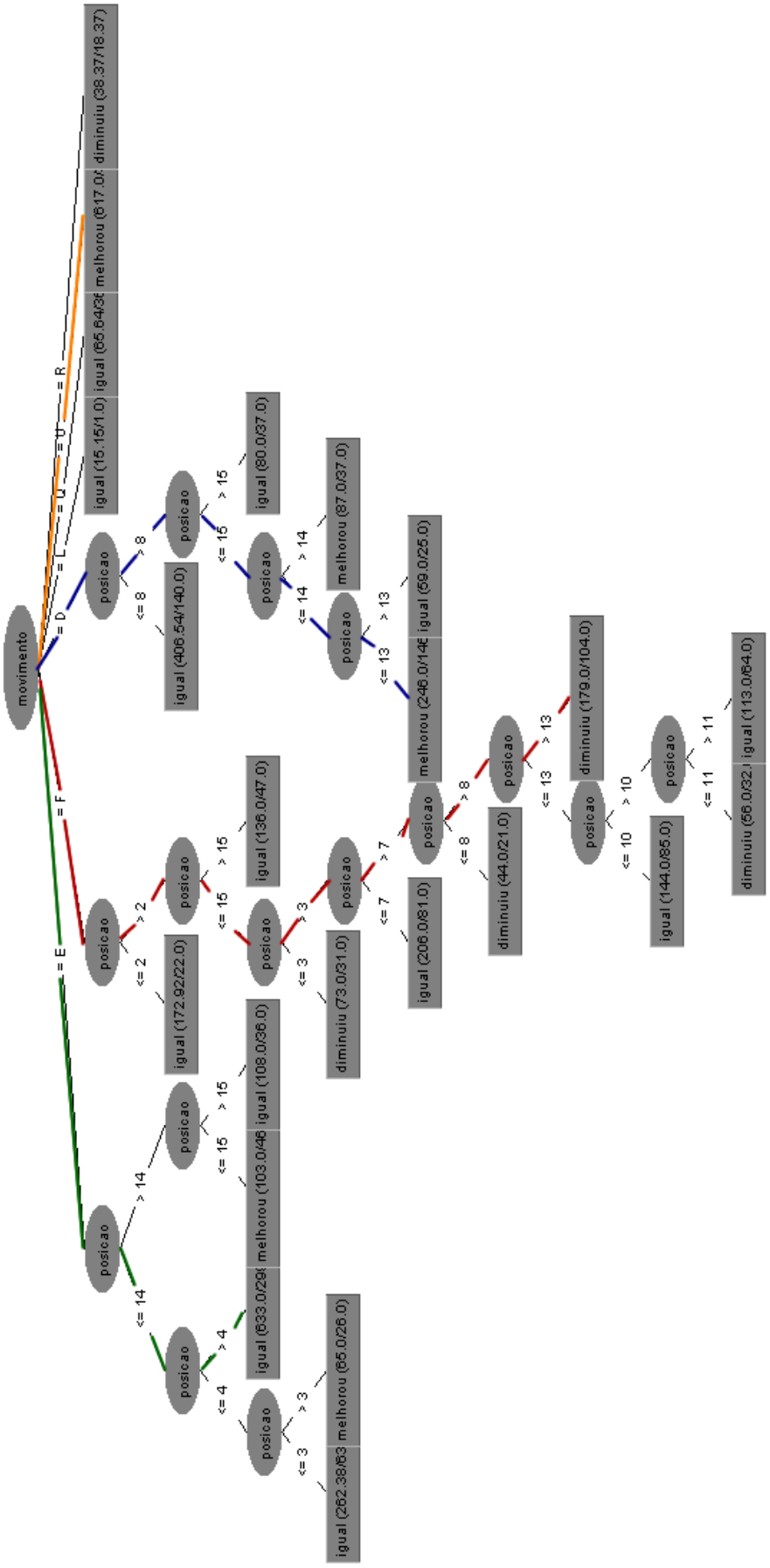
Caminho Vermelho: Este caminho mostra que com o movimento F(frente) nos aminoácidos 14 ou 15, a energia da estrutura diminui.

Caminho Azul: Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos entre 9 e 13 a energia da estrutura melhora.

Caminho Verde: Este caminho mostra que deixando E(esquerda) nos aminoácidos entre 5 e 14, a energia da estrutura permanece igual.

Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar).

Figura 32: Árvore de classificação gerada no teste 2 versão 2



Apesar de achar possíveis regras, este teste não teve uma taxa de acerto muito boa e pode-se observar que na matriz de confusão (Tabela 7) obteve uma grande desigualdade da melhoria igual para as outras assim como a versão 1, portanto os próximos testes irão buscar melhorar o percentual de acerto e a matriz de confusão.

4.4 Testes 3

- **Versão 1**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: jogada, movimento, posição e melhoria com um número mínimo de 19 instâncias por folha e seu atributo classificador foi o melhoria. Foram descartadas as jogadas de número zero e número um, assim como jogadas com melhoria = *igual*.

  Neste teste foi obtido uma árvore classificatória como mostra a Figura 33, que contém uma taxa de 67.47% de acerto, e uma matriz de confusão como mostra na Tabela 8.

  Tabela 8: Matriz de confusão teste 3 versão 1

	A	B
A = piorou	**489**	353
B = melhorou	278	**820**

  Conforme a Tabela 8, pode-se observar que:

  – As instâncias com melhoria = *melhorou* obteve 820 instâncias classificadas corretamente

  – As instâncias com melhoria = *piorou* obteve 489 instâncias classificadas corretamente

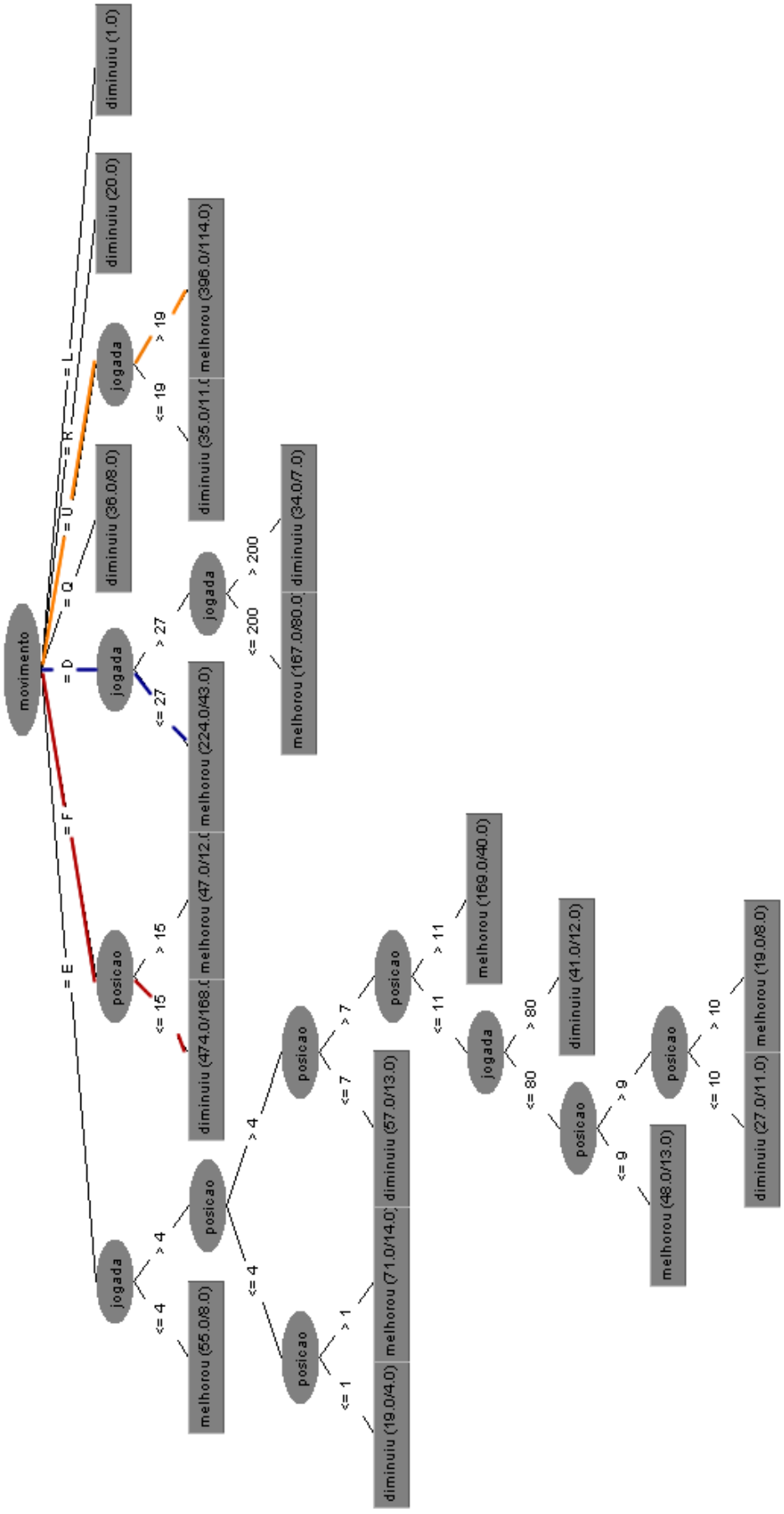
  Pode-se observar que na Figura 33 existem caminhos classificatórios que foram marcados com diferentes cores devido a relevância de seus números de instâncias classificadas:

  Caminho Vermelho: Este caminho mostra que com o movimento F(frente) nos aminoácidos 14 ou 15, a energia da estrutura diminui.

  Caminho Azul: Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos entre 9 e 13 a energia da estrutura melhora.

  Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar) em jogadas > 19.

Figura 33: Árvore de classificação gerada no teste 3 versão 1



Este teste obteve uma taxa de acerto razoavelmente boa e pode-se observar que a matriz de confusão (Tabela 8) obteve uma maior igualdade nas classificações melhorou e piorou.

- **Versão 2**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: movimento, posição e melhoria com um número mínimo de 19 instâncias por folha e seu atributo classificador foi melhoria. Foram descartadas as jogadas de número zero e número um, assim como jogadas com melhoria = *igual*.

  Neste teste foi obtido uma árvore classificatória como mostra a Figura 34 que contém uma taxa de 68.04% de acerto, e uma matriz de confusão como mostra na Tabela 9.

Tabela 9: Matriz de confusão teste 3 versão 2

	A	B
A = piorou	**441**	401
B = melhorou	219	**879**

Conforme a Tabela 9, pode-se observar que:

- As instâncias com melhoria = *melhorou* obteve 879 instâncias classificadas corretamente
- As instâncias com melhoria = *piorou* obteve 441 instâncias classificadas corretamente

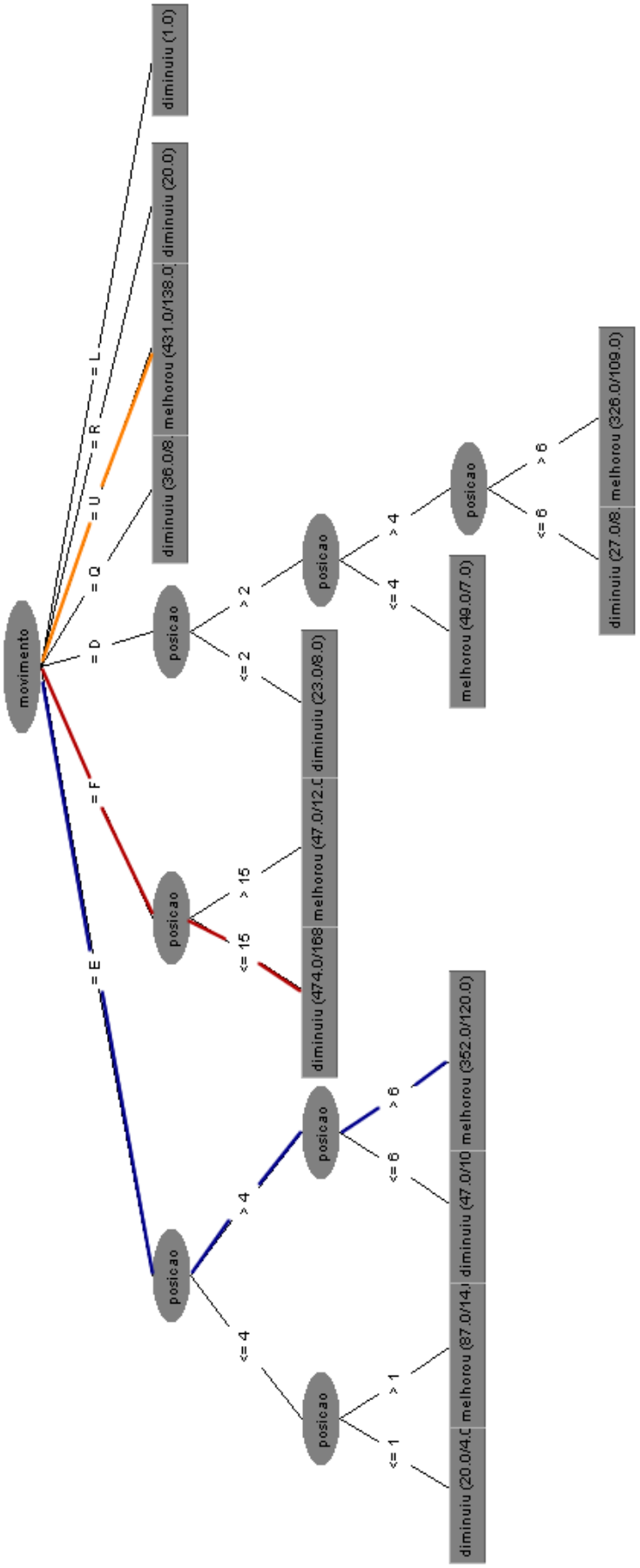
Pode-se observar que na Figura 34 existem caminhos classificatórios que foram marcados com diferentes cores, devido a relevância de seus números de instâncias classificadas:

Caminho Vermelho: Este caminho mostra que com o movimento F(frente) nos aminoácidos $>= 15$, a energia da estrutura diminui.

Caminho Azul: Este caminho mostra que efetuando um movimento E(esquerda) nos aminoácidos > 6 a energia da estrutura melhora.

Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar).

Figura 34: Árvore de classificação gerada no teste 3 versão 2



Este teste obteve uma taxa de acerto razoavelmente boa e pode-se observar que a matriz de confusão (Tabela 9) obteve uma maior igualdade das classificações melhorou e piorou.

4.5 Testes 4

- **Versão 1**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: jogada, movimento, posição e melhoria com um número mínimo de 29 instâncias por folha e seu atributo classificador foi melhoria. Foram descartadas as jogadas de número zero e número um, assim como 920 jogadas com melhoria = *igual* para balancear a quantidade de instâncias.

  Neste teste foi obtido uma árvore classificatória, como mostra a Figura 35 que contém uma taxa de 52.97% de acerto, e uma matriz de confusão como mostra na Tabela 10.

  Tabela 10: Matriz de confusão teste 4 versão 1

	A	B	C
A = piorou	**352**	270	220
B = igual	197	**574**	249
C = melhorou	164	292	**642**

  Conforme a Tabela 10, pode-se observar que:

  – As instâncias com melhoria = *melhorou* obteve 642 instâncias classificadas corretamente

  – As instâncias com melhoria = *piorou* obteve 352 instâncias classificadas corretamente

  – As instâncias com melhoria = *igual* obteve 574 instâncias classificadas corretamente

  Pode-se observar que na Figura 35 existem caminhos classificatórios que foram marcados com diferentes cores, devido a relevância de seus números de instâncias classificadas:

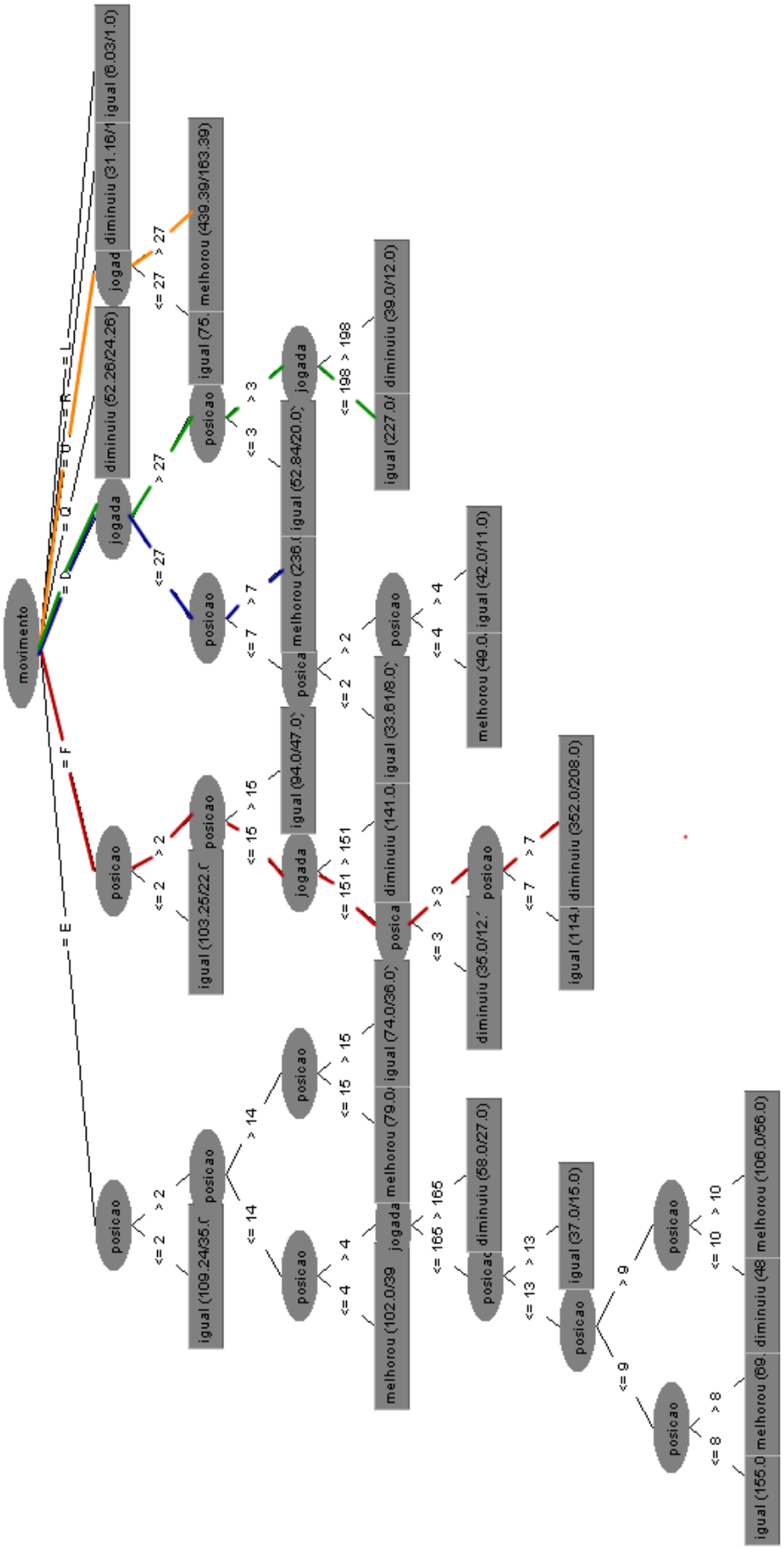
  Caminho Vermelho: Este caminho mostra que com o movimento F(frente) nos aminoácidos entre 4 e 6 nas jogadas >= 151, a energia da estrutura diminui.

  Caminho Azul: Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos > 7 nas jogadas <=27, a energia da estrutura melhora.

  Caminho Verde: Este caminho mostra que deixando D(direita) nos aminoácidos > 3 nas jogadas entre 28 e 198, a energia da estrutura permanece igual.

Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar) nas jogadas $>$27.

Figura 35: Árvore de classificação gerada no teste 4 versão 1



Este teste obteve uma taxa de acerto não muito boa, mas pode-se observar que a matriz de confusão (Tabela 10) obteve uma melhor igualdade das classificações melhorou e piorou.

- **Versão 2**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: movimento, posição e melhoria com um número mínimo de 29 instâncias por folha e seu atributo classificador foi o melhoria. Foram descartadas as jogadas de número zero e número um, assim como 920 jogadas com melhoria = *igual* para balancear a quantidade de instâncias.

  Neste teste foi obtido uma árvore classificatória como mostra a Figura 36 que contém uma taxa de 53.10% de acerto, e uma matriz de confusão como mostra na Tabela 11.

Tabela 11: Matriz de confusão teste 4 versão 2

	A	B	C
B = piorou	**366**	172	304
C = igual	203	**485**	332
A = melhorou	161	216	**721**

Conforme a Tabela 11, pode-se observar que:

- As instâncias com melhoria = *melhorou* obteve 721 instâncias classificadas corretamente
- As instâncias com melhoria = *piorou* obteve 366 instâncias classificadas corretamente
- As instâncias com melhoria = *igual* obteve 485 instâncias classificadas corretamente

Pode-se observar que na Figura 36 existem caminhos classificatórios que foram marcados com diferentes cores devido a relevância de seus números de instâncias classificadas:

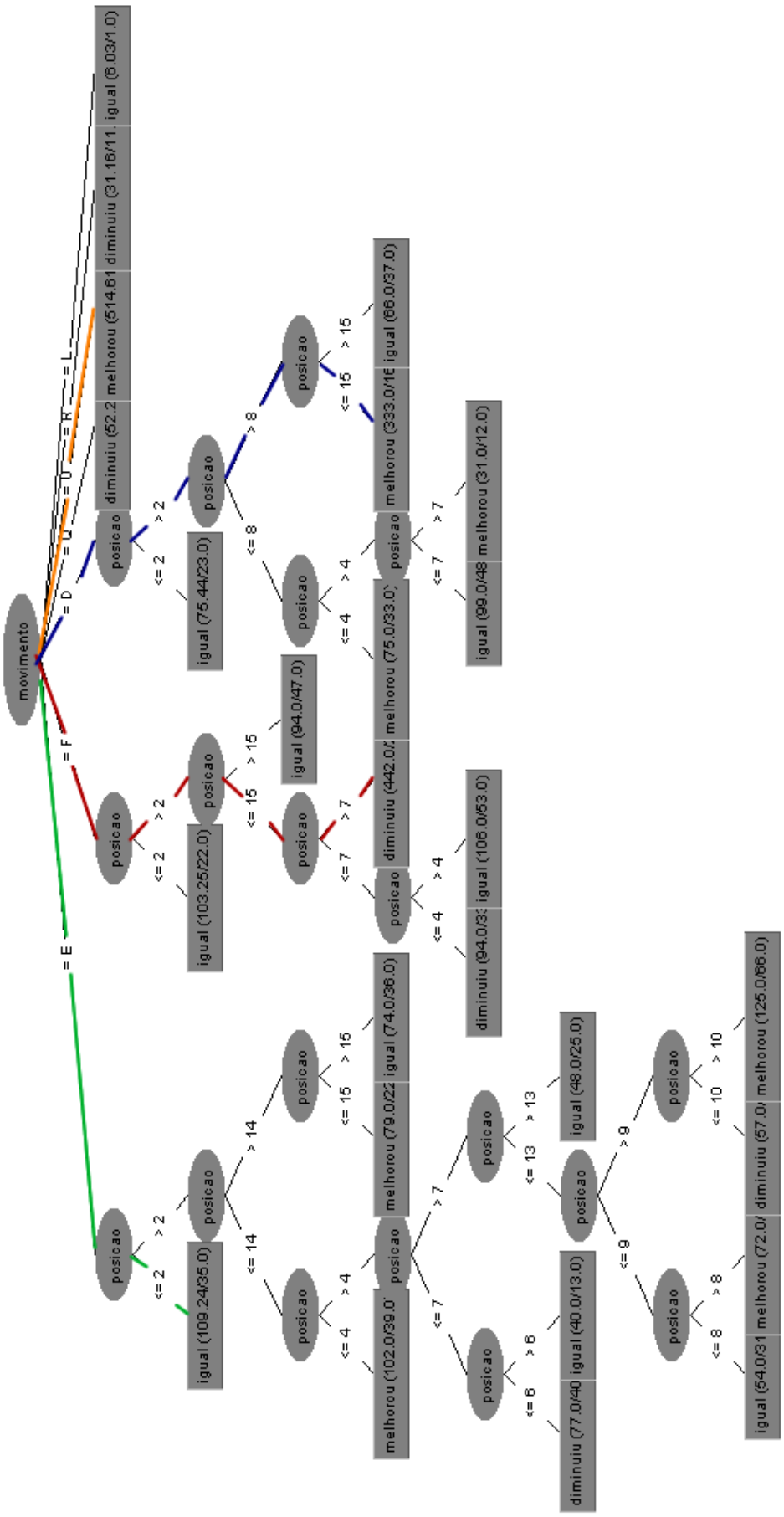
Caminho Vermelho: Este caminho mostra que com o movimento F(frente) nos aminoácidos entre 7 e 15, a energia da estrutura diminui.

Caminho Azul: Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos entre 9 e 15, a energia da estrutura melhora.

Caminho Verde: Este caminho mostra que deixando F(frente) nos aminoácidos $<= 2$, a energia da estrutura permanece igual.

Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar).

Figura 36: Árvore de classificação gerada no teste 4 versão 2



Este teste não obteve uma taxa de acerto boa, mas pode-se observar que a matriz de confusão (Tabela 11) obteve uma melhor igualdade das classificações melhorou e piorou.

4.6 Testes 5

- **Versão 1**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: jogada, movimento, posição e melhoria com um número mínimo de 19 instâncias por folha e seu atributo classificador foi melhoria. Foram descartadas as jogadas de número zero e número um, assim como jogadas com melhoria = *igual*.

  Neste teste foi obtido uma árvore classificatória como mostra a Figura 37 que contém uma taxa de 67.98% de acerto, e uma matriz de confusão como mostra na Tabela 12.

  Tabela 12: Matriz de confusão teste 5 versão 1

	A	B
A = igual	**489**	353
B = melhorou	278	**820**

  Conforme a Tabela 12, pode-se observar que:

  – As instâncias com melhoria = *igual* obteve 820 instâncias classificadas corretamente

  – As instâncias com melhoria = *igual* obteve 489 instâncias classificadas corretamente

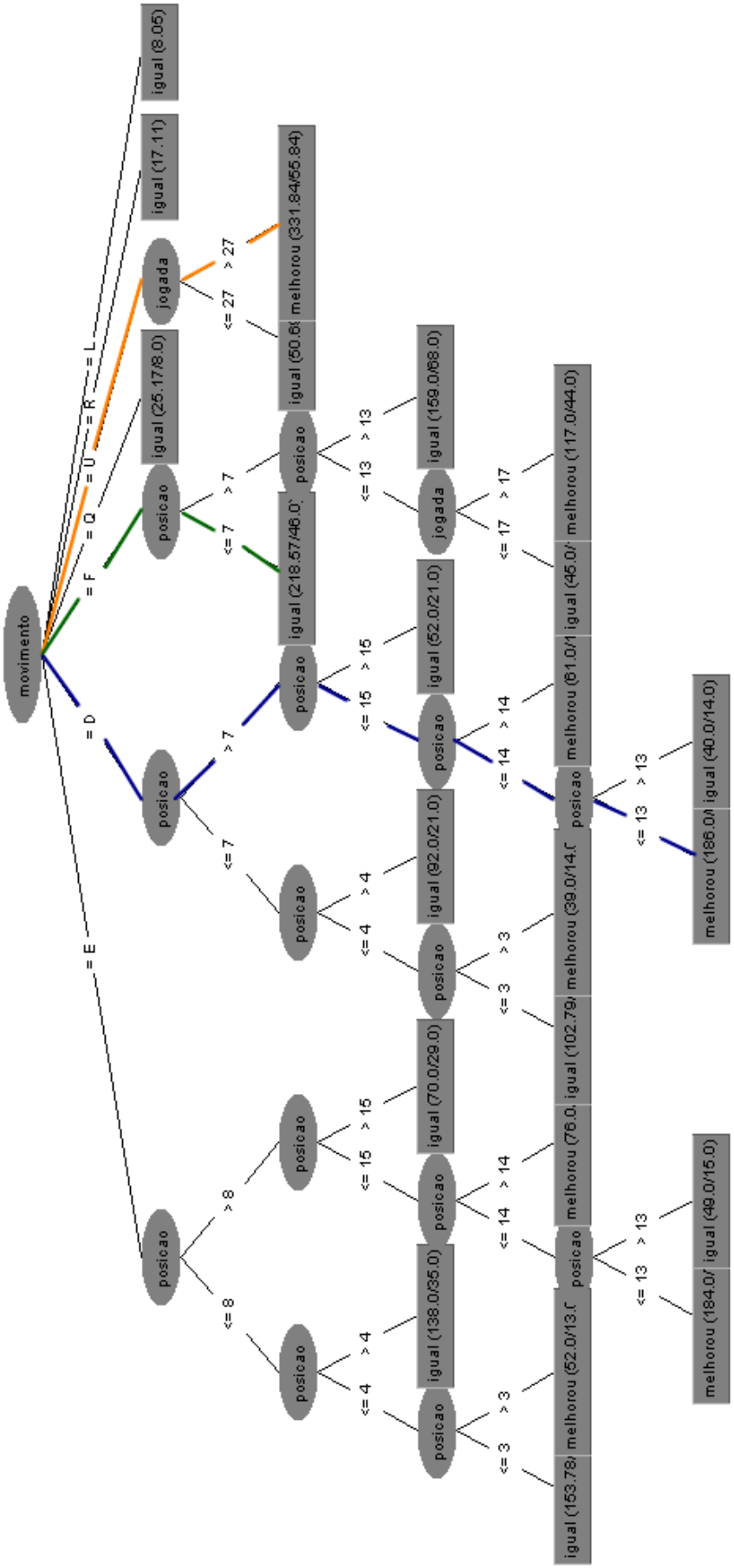
  Pode-se observar que na Figura 37 existem caminhos classificatórios que foram marcados com diferentes cores, devido a relevância de seus números de instâncias classificadas:

  Caminho Azul: Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos entre 8 e 13 a energia da estrutura melhora.

  Caminho Verde: Este caminho mostra que deixando F(frente) nos aminoácidos $<= 7$, a energia da estrutura permanece igual.

  Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar) nas jogadas > 27.

Figura 37: Árvore de classificação gerada no teste 5 versão 1



Neste teste, apesar do baixo número de instâncias obteve-se uma taxa de acerto razoavelmente boa, e pode-se observar que a matriz de confusão (Tabela 12) obteve uma maior igualdade das classificações melhorou e igual.

- **Versão 2**

  As instâncias fornecidas neste teste foram classificadas contendo os atributos: movimento, posição e melhoria com um número mínimo de 19 instâncias por folha e seu atributo classificador foi melhoria. Foram descartadas as jogadas de número zero e número um, assim como jogadas com melhoria = *igual*.

  Neste teste foi obtido uma árvore classificatória como mostra a Figura 38, que contém uma taxa de 67.63% de acerto, e uma matriz de confusão como mostra na Tabela 13.

Tabela 13: Matriz de confusão teste 5 versão 2

	A	B
A = igual	**838**	332
B = melhorou	394	**704**

Conforme a Tabela 13, pode-se observar que:

  – As instâncias com melhoria = *igual* obteve 704 instâncias classificadas corretamente

  – As instâncias com melhoria = *igual* obteve 838 instâncias classificadas corretamente

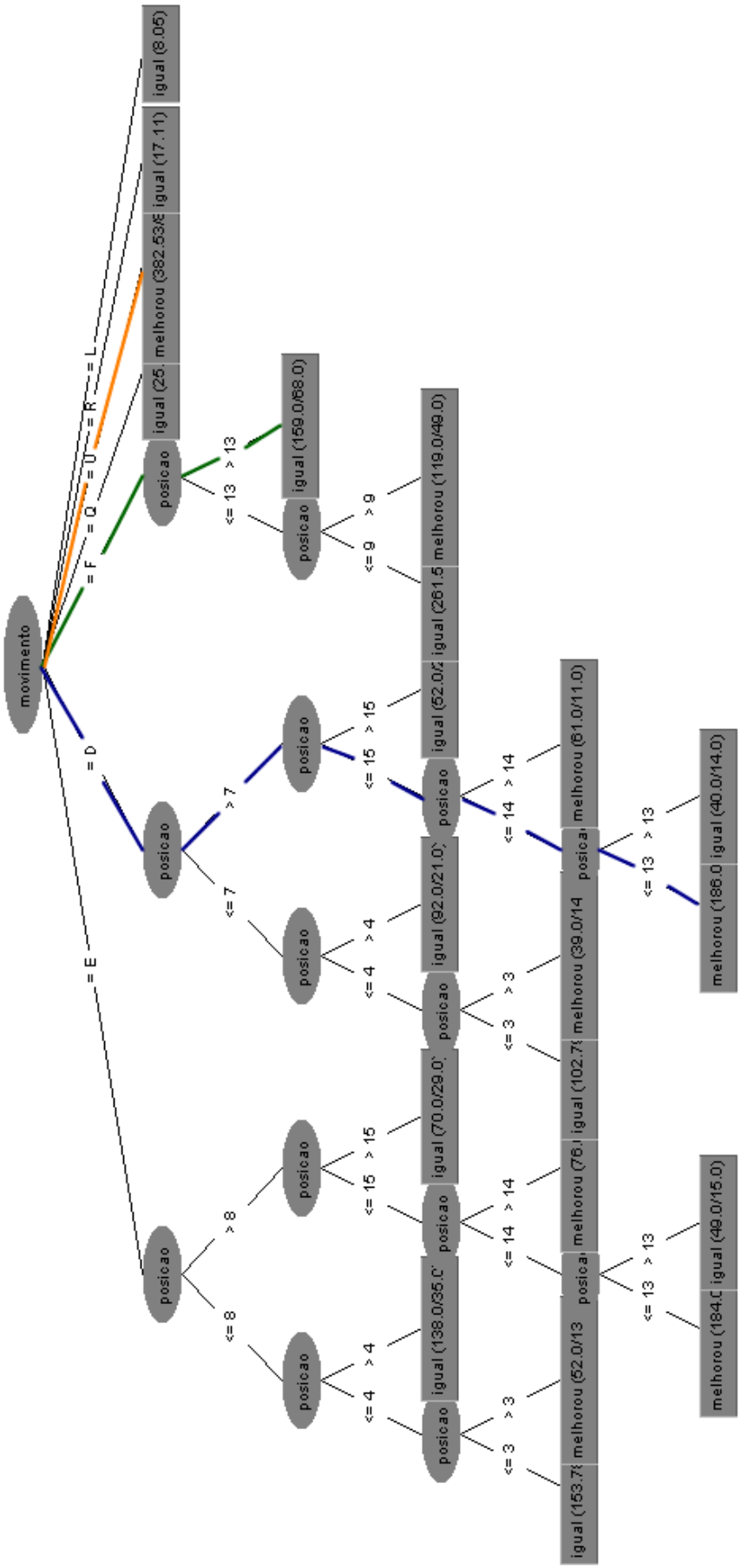
Pode-se observar que na Figura 38 existem caminhos classificatórios que foram marcados com diferentes cores devido a relevância de seus números de instâncias classificadas:

Caminho Azul: Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos entre 8 e 13 a energia da estrutura melhora.

Caminho Verde: Este caminho mostra que deixando F(frente) nos aminoácidos < 13, a energia da estrutura permanece igual.

Caminho Laranja: Este caminho mostra que a energia melhora quando o jogador utiliza a ferramenta U(voltar).

Figura 38: Árvore de classificação gerada no teste 5 versão 2



Neste teste, apesar do baixo número de instâncias, obteve uma taxa de acerto razoavelmente boa e pode-se observar que a matriz de confusão (Tabela 13) obteve uma maior igualdade das classificações melhorou e igual.

4.7 Considerações finais

Com os resultados adquiridos nos 5 testes foram obtidas algumas estratégias de dobramento de proteína, como as do teste 3 versão 2, que obteve um percentual de acerto mais elevado que a dos outros testes. Este teste mostra que:

- Com o movimento F(frente) nos aminoácidos $>= 15$, a energia da estrutura diminui.
- Efetuando um movimento E(esquerda) nos aminoácidos > 6 a energia da estrutura melhora.
- A energia melhora quando o jogador utiliza a ferramenta U(voltar).

É possível analisar que não apenas no teste 3 versão 2 e sim em todos os outros testes, obteve-se melhora da energia ao utilizar a ferramenta U (voltar) no jogo. A explicação disso é que há uma grande tendência dos jogadores utilizarem essa ferramenta logo após efetuar um dobramento onde piora a energia, sendo assim, voltando a posição anterior onde a estrutura encontra-se em uma posição com melhores resultados.

Foi observado também que, por algum motivo, as classificações melhoria $=$ *piorou* contém muitas classificações erradas como mostram as matrizes de confusão dos testes. E nos testes 5 onde foram ignoradas as melhorias $=$ piorou, obteve uma melhor porcentagem de acerto, assim como nos testes onde foram ignoradas as jogadas de melhoria $=$ *igual* devido ao grande desbalanceamento de instâncias observada nas matrizes de confusão. Os testes não obtiveram melhora quando foram ignoradas apenas algumas instâncias de melhoria $=$ *igual*, com o objetivo de balancear o número de instâncias.

5 Conclusões e Trabalhos Futuros

Existem muitos comportamentos ainda não desvendados a respeito das proteínas, e sempre será importante buscar respostas de perguntas ainda não respondidas sobre este assunto. Devido às limitações tecnológicas que não conseguem desvendar alguns destes problemas, muitos cientistas estão em busca de conhecimentos sobre o assunto, como buscar extrair conhecimentos sobre dobramento de proteína com a ajuda da inteligência humana, onde já existem resultados positivos dessa prática.

Analisando os resultados obtidos através dos testes realizados neste projeto, chega-se a conclusão de que é realmente possível obter e extrair conhecimentos de uma base de dados de um jogo sério de dobramento de proteína, utilizando técnicas de mineração de dados, onde todas as estratégias obtidas foram através da inteligência humana.

Percebe-se também que os resultados obtidos além de obter algumas estratégias de dobramento, serviram como um pré-processamento de dados e sendo assim descobrindo outras formas de se obter novos e mais relevantes conhecimentos através da inteligência humana.

A mineração de dados mostrou-se capaz de ajudar na descoberta dos conhecimentos, mas ainda existe uma vasta gama de algoritmos a serem explorados e submetidos a testes.

Desta forma, como trabalhos futuros, este projeto pode visar obter um maior volume de dados em sua base de dados; melhorar as ferramentas fornecidas para os jogadores, facilitando a jogabilidade; e tornar o jogo mais dinâmico e *multiplayer*(capaz de deixar 2 ou mais jogadores interagirem juntos na mesma partida em tempo real); e após obter boas estratégias de dobramento, uma etapa subsequente ao projeto, será criado um jogador artificial, levando a obter possíveis dobramentos que ainda não foram descobertos.

Devido a complexidade e a interdisciplinaridade do assunto abordado neste projeto, seria interessante agregar a estes trabalhos futuros, profissionais e pesquisadores de áreas como: interfaces Humano, estatística e biologia, transformando em um projeto mais robusto e conceituado.

Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. *Proc. of the ACM SIGMOD*, v. 97, p. 207–216, 1993. Citado na página 28.

ALLé, J. M. *O Grande Livro dos Jogos*. [S.l.]: Leitura, 1999. Citado na página 24.

BAKER, D. A surprising simplicity to protein folding. *Nature*, v. 405, n. 6782, p. 39–42, 2000. Citado 5 vezes nas páginas 6, 20, 23, 24 e 25.

BRANDEN, C.; TOOZE, J. *Introduction to Protein Structure*. [S.l.]: Garland Science, 1999. Citado 7 vezes nas páginas 6, 15, 16, 17, 18, 19 e 20.

CABENA, P. et al. *Discovering Data Mining: From Concept to Implementation*. [S.l.]: Prentice Hall, 1998. Citado na página 27.

CHANDRU, V.; DATTASHARMA, A.; KUMAR, V. A. The algorithmics of folding proteins on lattices. *Discrete Applied Mathematics*, v. 405, n. 127, p. 145–161, 2003. Citado na página 20.

CIOS, K. J. et al. *Data Mining – A Knowledge Discovery Approach*. [S.l.]: Springer, 2007. Citado na página 26.

COOPER, S. et al. Predicting protein structures with a multiplayer online game. *Nature*, v. 466, p. 756–760, 2010a. Citado na página 12.

COOPER, S. et al. The challenge of designing scientific discovery games. *Foundations of Digital Games*, p. 40–47, 2010b. Citado 6 vezes nas páginas 6, 12, 24, 30, 31 e 32.

DILL, K. et al. Principles of protein folding – a perspective from simple exact models. *Protein science*, v. 4, p. 561–602, 1995. Citado 3 vezes nas páginas 21, 22 e 23.

DILL, K. A. Theory for the folding and stability of globular proteinst. *Biochemistry*, v. 24, p. 1501–1509, 1985. Citado 5 vezes nas páginas 6, 12, 21, 22 e 23.

DINNER, A. et al. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences*, v. 25, p. 331–339, 2000. Citado na página 20.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From Data Mining to Knowledge Discovery in Databases*. [S.l.]: American Association for Artificial Intelligence, 1996. Citado 3 vezes nas páginas 6, 26 e 27.

FIALHO, N. *Jogos no Ensino de Química e Biologia*. [S.l.]: IBPEX, 2007. Citado na página 24.

FOLDIT. *Foldit*. 2012. Foldit Web Site. Disponível em: <www.fold.it/portal/>. Acesso em: 21.4.2012. Citado na página 31.

GARCIA, R. *Jogo on-line ajuda cientistas a fazer mapa do cérebro*. 2013. Folha de S. Paulo. Disponível em: <www1.folha.uol.com.br/ciencia/2013/05/1274984-jogo-on-line-ajuda-cientistas-a-fazer-mapa-do-cerebro.shtml>. Acesso em: 10.6.2012. Citado 4 vezes nas páginas 6, 32, 33 e 34.

HALL, M. et al. *The weka data mining software: An update*. [S.l.]: SIGKDD Explorations, 2009. Citado na página 47.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Elsevier, 2006. Citado 3 vezes nas páginas 26, 28 e 29.

HART, W.; ISTRAIL, S. *Hp benchmarks*. 2012. Sandia Web Site. Disponível em: <www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html>. Acesso em: 23.3.2013. Citado 3 vezes nas páginas 8, 21 e 22.

HUIZINGA, J. *Homo Ludens: A Study of the Play-Element in Culture*. [S.l.]: Beacon Press, 1971. Citado na página 23.

JAIN, V.; SEUNG, H. S.; TURAGA, S. C. *Machines that learn to segment images: a crucial technology for connectomics*. [S.l.]: Neurobiology, 2010. Citado na página 34.

KHATIB, F. et al. Mining association rules between sets of items in large databases. *Nature Structural & Molecular Biology*, v. 18, p. 1175–1177, 2011. Citado na página 31.

LAROSE, D. T. *Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons*. [S.l.]: John Wiley and Sons, Inc, 2005. Citado na página 26.

LEHNINGER; NELSON, D. L.; COX, M. M. *Lehninger Principles of Biochemistry*. [S.l.]: W. H. Freeman, 2008. Citado 3 vezes nas páginas 8, 15 e 16.

LOULA, A. C. et al. Modelagem ambiental em um jogo eletrônico educativo. *VIII Brazilian Symposium on Games and Digital Entertainment*, 2011. Citado 2 vezes nas páginas 6 e 35.

MIRANDA, M. J. A inteligência humana: contornos da pesquisa. *Paidéia (Ribeirão Preto)*, v. 12, n. 23, p. 19–29, 2002. Citado na página 12.

NIELSEN, J. *Designing Web Usability: The Practice of Simplicity*. [S.l.]: New Riders, 2000. Citado na página 25.

PLOTKIN, S.; ONUCHIC, J. Investigation of routes and funnels in protein folding by free energy functional methods. *Proceedings of the National Academy of Sciences*, v. 97, p. 6509–6514, 2000. Citado na página 20.

PRENSKY, M. *Digital Game-Based Learning*. [S.l.]: McGraw-Hill Pub. Co., 2004. Citado na página 24.

PTITSYN, O. B. A determinable but unresolved problem. *The FASEB Journal*, v. 10, p. 3–4, 1996. Citado na página 12.

RODRIGUES, A. F.; AMARAL, L. R. Aplicação de métodos computacionais de mineração de dados na classificação e seleção de oncogenes medidos por microarray. *Revista Brasileira de Cancerologia*, v. 58, p. 241–249, 2012. Citado 2 vezes nas páginas 35 e 36.

TROVATO, A. et al. What determines the structures of native folds of proteins? *Journal of Physics: Condensed Matter*, v. 17, p. 1515–1522, 2005. Citado 3 vezes nas páginas 15, 19 e 20.